Complexity

Artificial Intelligence for Smart System Simulation

Lead Guest Editor: Aboul Ella Hassanien
Guest Editors: Hye-jin Kim, Borut Buchmeister, and Subramaniam Ganesan

Artificial Intelligence for Smart System Simulation

Complexity

Artificial Intelligence for Smart System Simulation

Lead Guest Editor: Aboul Ella Hassanien
Guest Editors: Hye-jin Kim, Borut Buchmeister, and Subramaniam Ganesan

Contents

Contents

Hindawi
Complexity
Volume 2021, Article ID 5541449, 12 pages
https://doi.org/10.1155/2021/5541449

Research Article

Energy-Efficient Clustering and Localization Technique Using Genetic Algorithm in Wireless Sensor Networks

Junfeng Chen, Samson Hansen Sackey, Joseph Henry Anajemba, Xuewu Zhang, and Yurun He

College of Internet of Things Engineering, Hohai University, Changzhou 213022, China

Correspondence should be addressed to Junfeng Chen; chen-1997@163.com

Received 18 February 2021; Accepted 20 July 2021; Published 3 August 2021

Academic Editor: Abd E.I.-Baset Hassanien

Localization is recognized among the topmost vital features in numerous wireless sensor network (WSN) applications. This paper puts forward energy-efficient clustering and localization centered on genetic algorithm (ECGAL), in which the residual energy, distance estimation, and coverage connection are developed to form the fitness function. This function is certainly fast to run. The proposed ECGAL exhausts a lesser amount of energy and extends wireless network existence. Finally, the simulations are carried out to assess the performance of the proposed algorithm. Experimental results show that the proposed algorithm approximates the unknown node location and provides minimum localization error.

1. Introduction

The application of wireless sensor networks for precise localization of devices and humans in bounded scenarios characterizes a significant feature in the provision of highly classified essential services in the discipline of logistics and administration [1]. The basic implementation force for various localization means is monitoring, sensing, and tracking. However, few instances include health monitoring, habitat monitoring, agriculture monitoring, environment monitoring, military investigation, intruder recognition, pollution control, and space handling on planet and moon. WSN is perceived to be in any environmental structure such as underwater and terrestrial nature. Presently, the persistent distribution of such systems is prohibited as a result of their gelatinous intricacy and their costs of maintenance [2]. Primarily, this is owing to the fact that, in interior environments, precise localization in the occurrence of unembellished multipath fading frequently necessitates a huge amount of static sensing nodes which are referred to as anchors (also called beacon) with a recognized location [3] and the utilization of complex signal processing algorithms for disseminated proximity approximation [4]. In a broader spectrum, accumulating more anchor nodes means higher position accuracy. However, anchor nodes cost more than ordinary nodes and are 10 times more expensive [5]. By and large, if the position of an unidentified node is discovered, the anchor nodes will then be negligibly ignored. Consequently, with the decrease of quantity of anchor nodes, the positioning of the node is affected [6]. Reducing anchor nodes cuts cost, and the more the anchor nodes are slashed, the lesser the accuracy is. Lately, the approach used in WSN localization is based on the optimization problem of multidimensional and multimodal which is solvable by utilizing stochastic methods which are based on population [7]. Section 2 of our research reviews several metaheuristic algorithms used in the field of WSN localization. The results of those research studies show that these algorithms have successfully and drastically minimized errors in localization [8].

The two distinctive ways by which localization can be performed are distributed and centralized localization. In the former, each node finds the unknown nodes by themselves, while in the latter, information from each node is transmitted to a centralized component for further processing in order to obtain information about the position. Furthermore, the clustering process of the nodes in the second one is seen to be divided into the network that is interconnected, called clusters, with each cluster having

many energized sensor nodes led by a cluster center (cluster head) node which is a coordinator. The cluster center works in sequence with other cluster centers which are all provisional base stations. Most WSN nodes are categorized into several states including isolated, normal, cluster center, and gateway. Ultimately, isolated nodes preserve a secure neighbor table where neighbor node information is stored. The selection of cluster centers is considered very essential in clustering. Clustering in sensor networks is to a large extent power efficient. In a broader term, localization methods are categorized into range-free and range-based methods [9]. Range-based localization exploits the distance connecting an unidentified node and a recognized node (anchor) even though range-free methods solely utilize nodes' connectivity data. On the other hand, range-based methods are fine-grained localization methods, whereas range-free methods are coarse-grained [10]. Range-based methods can be classified into four major classes, namely, time-difference-of-arrival (TDOA), time-of-arrival (TOA), angle-of-arrival (AOA), and received signal strength (RSS). On the other hand, range-free localization methods include centroid algorithm, Monte Carlo localization, approximate point in triangle test (APTT), closest point-based method, assumption-based coordinate method, DV-HOP, and amorphous method.

Energized sensor nodes in hostile environments must stay alive for long time, but it is ambiguous or almost impossible to boost or change their batteries [11], and this necessitates inventing new energy-efficient alternatives to some of the existing traditional wireless networking challenges which include intermediate access control, self-organization, bandwidth distribution, security, and routing. Some factors to consider in prolonging the lifetime of networks is by utilizing the gains of trade-offs such as power, latency, and accuracy, coupled with using hierarchical (tiered) architectures. The position of a localized unknown node is important because it helps solve bigger WSN problems like routing and data aggregation. For this reason, it is necessary to focus on localization methods in WSN works [12]. Localization occurs once when considering static nodes, whereas tracking is nonstop localization of the cell node over time. Clustering is a preferred method for attaining competent and accessible overall performance in WSNs [13]. Clustering nodes into sets conserves power and minimizes disputes at the network since the nodes transmit their information to their cluster heads through reduced distances. Thus, the circulation of command through the transmission facilitates permits locality of transmission [14]. With respect to the difficulties faced by energized sensor nodes in localization, this paper's aim is to establish an elevated energy-efficient localization technique which is dependent on low energy depletion and a clustering structure. The proposed approach continues to show strength by dealing with location accuracy via clustering means in GA localization. Additionally, ECGAL successively reduces the whole WSN energy consumption.

In summary, the main contributions of this paper are summarized as follows:

(1). We utilized genetic algorithm (GA) with an energy-efficient clustering approach to solving localization problems in WSN.

(2) The performance of the proposed ECGAL (energy-efficient clustering in genetic algorithm localization) is analyzed and compared with DV-HOP (distance vector-hop), CENTA (centroid algorithm), EDV-HOP (evolutionary distance vector-hop), and CGAL (clustering in genetic algorithm localization).

(3) The results have shown that our proposed approach outperforms the existing localization algorithms with respect to energy efficiency, localized distance error, and coverage connection.

The remaining part of this paper is structured as follows. Section 2 elaborates several previous research exertions relating to localization. In Section 3, a brief description of the utilized genetic algorithm (GA) is presented, while Section 4 analyzes a sensor node localization setup and cluster formation in a wireless sensor network localization scenario. Section 5 provides details on the fitness function proposal, including the definitions of energy efficiency, distance estimation, and coverage connection. Section 6 presents the performance evaluation of ECGAL. Finally, the research summary and derived conclusions are presented in Section 7.

2. Related Works

Lately, there are several algorithms in the field of optimization used in reproving the drawbacks in the localization of WSN nodes [15]. Some of these existing related research studies are briefly discussed in this section.

The authors in [16] proposed an efficient hybrid bio-inspired optimization in localization methods which is applicable in industrial WSNs. In their research, they proposed particle swarm optimization (PSO) and dragonfly algorithm (DA) which earns slight time of computation and extensive accuracy. On the other hand, Kanoosh et al. suggested a node localization design which is dependent on a current bio-inspired algorithm known as salp swarm algorithm (SSA). The performance of their results is measured against similar optimization algorithms, specifically, particle swarm optimization (PSO), firefly algorithm (FA), grey wolf optimizer (GWO), and butterfly optimization algorithm (BOA) under distinctive wireless sensor network positioning [17]. Based on the current localization and monitoring setups, the localization of sensor nodes and the application of WSN target trailing technology had been examined from the perspectives of accuracy perfection, extending the WSN natural life with respect to coordination theory, particle filter, range-free theory, and different computing approaches. Zhang et al. merged an energized sensor triggering algorithm and dynamic clustering process to prolong a parallel particle filter algorithm and further target monitoring system [18]. The authors adopted a two-object tracking strategy used in WSNs primarily based on cluster algorithms which have been combined together to perform many features in the proposed algorithm. Musafer et al.

benefited from using cluster algorithms to count and detect node in a cluster by reporting an event to the cluster center (also cluster head) node according to a query, conveying all audible information to the base station [19].

Numerous device-free localization methods are launched in wireless systems. Alippi et al. proposed a radio tomographic imaging- (RTI-) based localization technique imaging the received signal strength (RSS) reduction resulting from the targets with inexpensive and standard hardware [20, 21]. Xu et al. employed device-free wireless localization in WSNs which utilizes the RSS differences between sensor nodes in order to find solutions to problem [22, 23]. Wang et al. explained that when a target is found in a WSN deployment area, the existence of the target will imitate, strew, and engage the WSNs radio signals [24]. The outcomes of localization are determined with several RSS values using (3). Zhang and Wong exercised the genetic algorithm for gathering ecological questions within a WSN for successfully localizing its sensor nodes [25]. However, all the coordinates of the grid network offer random perturbations of the quality of the received signal. Furthermore, genetic algorithm can enormously acquire information about the location and minimize the likely errors related with the RSSI estimation assumed for every coordinate. Sackey et al. showed the implementation of GA to practically localize WSN nodes by means of three coordinates or more anchors [26].

The research of Wang et al. [27] centered on range-free localization as a cheap choice in comparison to range-based methods. But the localization based on range-free undergoes greater localization mistakes in contrast to the range-based algorithms. Furthermore, Sivakumar et al. offered an expanded form of DV-HOP which is a famous range-free method that is reliant on hop-proximity calculation [28]. Primarily, the enhancement in the DV-HOP algorithm is done depending on GA. Song et al. provided simulation results to prove the superior precision performance of the proposed localization algorithm in localization against the performance of other localization algorithms in [29]. Kumar discussed the localization of sensors in motion to deliver the advancement of connectivity, security, and energy consumption [30]. This needs to be an uncomplicated, dynamic, and adaptable localization method. Wang et al. presented a mobile sensor localization algorithm which is independent and has the capacity of coping with the ambiguities, connection breakdowns, and node flexibility in the network [31]. Nonetheless, computational intelligence (CI) possess qualities identical to such algorithms. Sharma et al. showed that CI methods can function in a setting of imprecision and ambiguity [32, 33]. CI consists of methods that can be altered. These methods can act logically in composite situations.

3. Brief Introduction of GA Algorithm

The proposed methodology relies on efficient clustering and strong global search for GA in order to increase accuracy and efficiency. Subsequently, the ensuing subsection presents the GA.

3.1. Genetic Algorithm. Genetic algorithm (GA) is derived from biological behaviors used in the field of optimization. GA is an existing metaheuristic driven by the approach of evolutional algorithms for natural selection. The assumed population size is said to consist of N^p competitors (candidates) on the possible solution which is made up of decoding and encoding chromosomes to a fixed dimension of binary numbers. The given interval contains 0 s and 1 s with N bits. The GA approach follows a specific procedure which is apportioned along with the genetic algorithm operators into initialization, selection, crossover, and mutation. Figure 1 shows a sample structure of a chromosome.

3.1.1. Initialization. Firstly, the countless mixed-up candidate solutions created tend to shape the preliminary population. The general population range is subjected to natural adversities but characteristically comprises a number of hundreds or more viable solutions. Conventionally, the now scattered populace produces an overlay which completes the varying feasible options (the search space). The search space entails all likely options to the question. Seldomly, the options might be "seeded" in the search space with high-quality solutions.

3.1.2. Selection. The technique of decision making relies on the chromosomes' fitness capacities to control the mating process for every individual. The chromosome holds the results in the shape of genes and is chosen in accordance to a particular selection method. The better the fitness value of these chromosomes, the greater the probability of being selected. Solutions with higher fitness have more chance to duplicate. However, ranking takes place after the chromosome with the highest fitness value attains the most appropriate chromosome. The defined function portrays the nearness a solution can change to, providing best results. Based on the value of likelihood selection, one or more individuals multiply to bring forth offspring. The probability of selecting is P^a of which each individual is determined by

$$P^a = \frac{\text{Fitness}^a}{\sum_{b=1}^{N^p} \text{Fitness}^b}, \tag{1}$$

where $a \in \{1, \ldots, N^p\}$, and the fitness of the selected individual a^{th} is denoted as Fitness^a. The selection of a chromosome denoted as a is based on $r \in (0, 1)$ random numbers. The cumulative probability C^a is well defined in (2), and it satisfies a chromosome selected at random within $C^{i-1} < r \leq C^a$.

3.1.3. Crossover. Crossover operator pairs two formerly selected chromosomes to copulate and produce offspring that share positive characteristics of both parents. Copulating comprises choosing two arbitrary crossover points c^1 and c^2 along the stretch of the chromosome. As a result, the encoded binary numbers are surrounded by some points that can be swapped between carefully chosen chromosomes interchangeably.

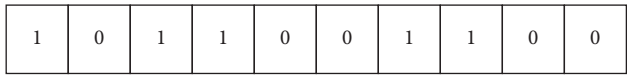


Figure 1: Structure of a sample chromosome.

$$C^a = \sum_{b=1}^{N^p} \text{Fitness}^b. \quad (2)$$

(1) Single-Point Crossover. In single-point crossovers, two parents can produce a cut point and re-merge the first fragment of the original parent crosses with the second fragment of the subsequent parent to create one offspring. The second fragment of the original parent is then fused with the first fragment of the subsequent parent to create another offspring. In Figure 2, a random point is preferred from two parents. Parents are then divided at the crossover point. Finally, children are created by exchanging tails.

(2) Two-Point Crossover. The single-point crossover and the two-point crossover share many similarities exempting the number of cut points they both create. In the two-point crossover, two offspring are created, but in the single-point crossover, only one cut point is made. As observed from the representation in Figure 3, there are two arbitrary numbers differing between 0 and 1 (length of the chromosome).The genes found between these two arbitrary chosen numbers are delivered from the initial parent to offspring and the complementary genes in the second parent are neglected. The vacant cells of the offspring are then singly filled with the unused genes from the second parent.

(3) N-Point Crossover. In Figure 4, n random crossover points are chosen from a chromosome sample, which is then fragmented along with those points. Then, exclusive parts are joined, which are alternating between parents. Finally, there is generality of 1 point (still some point preference).

3.1.4. Mutation. Immature concurrences in the algorithm are avoided since the mutation follows a GA mechanism which brings out uncharted results in the GA population. The arbitrary binary changes in chromosomes direct the process of mutation. The sequence from selection, mutation, and crossover is looped. The global optimum of a perfect but average individual is closely realized after multiple consecutive iterations of an expanding population.

(1) Displacement Mutation. The displacement mutation process arbitrarily selects two genes and transposes them after the parents are chosen. Figure 5 illustrates the implementation of the displacement mutation.

(2) Shift Mutation. Following the selection of the parent chromosome, two different points are carefully chosen randomly in a 1 to n interval (chromosome span) and the genetic factors positioned amidst these two points are moved towards the left corner, rotationally. An illustration of this shift mutation is presented in Figure 6.

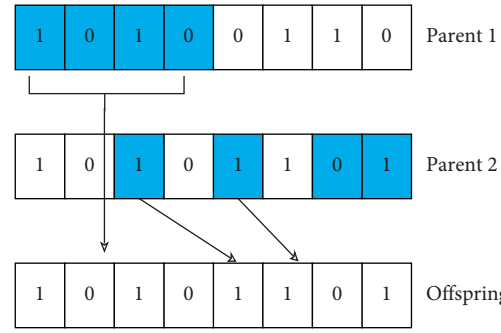


Figure 2: Single-point crossover.

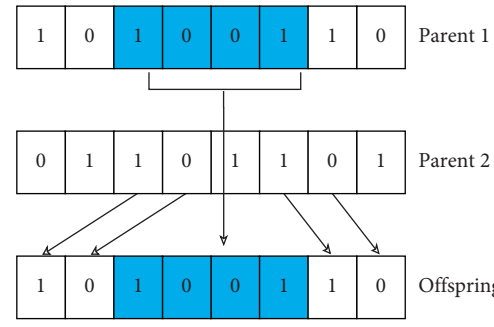


Figure 3: Two-point crossover.

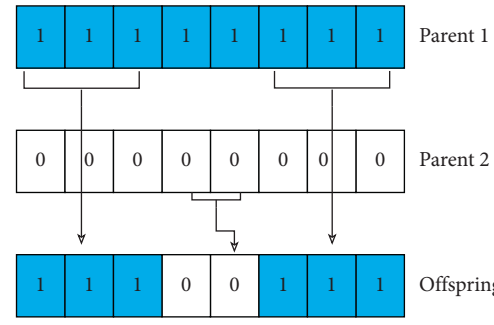


Figure 4: N-point crossover.

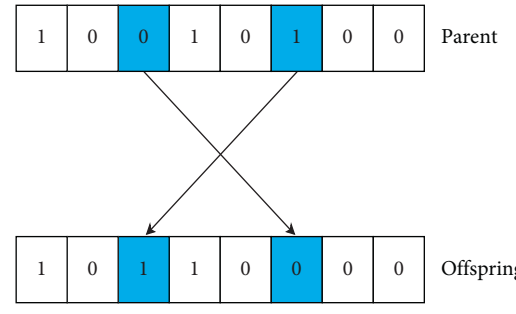


Figure 5: Displacement mutation.

3.1.5. Termination. The process of generation is recurrent until an end state is attained. Specifically, the termination criteria include the following: the value of the objective function reaches a certain predefined value, the number of iterations reaches the preset maximum iteration, the time or calculation cost of the budget allocation is reached, the

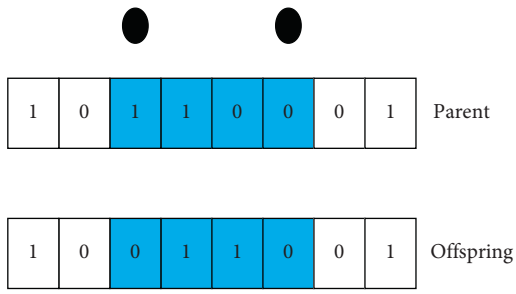


Figure 6: Shift mutation.

objective function value does not improve within a certain number of iterations, and various mixed termination criteria of the above termination conditions.

3.2. GA Pseudocode. The GA process is concisely analyzed in this section. Firstly, Figure 7 demonstrates the GA flow chart which describes the step-by-step approach of the proposed genetic algorithm. Secondly, the pseudocode for the GA is shown in Algorithm 1.

4. Design of Localization Problem for WSN

4.1. Description of Localization. Since the received signal strength (RSS) measurements are comparatively low-priced and simple to implement in hardware, they are widely used in real-world localization applications and are also approved as a preserved confined measurement in most research projects. The RSS at a reliable place within a localization area can be stated as follows:

$$\mathrm{RSS}(d_o) = P_t + K_e - 10\eta \log\left(\frac{d_o}{d_1}\right) + \alpha + \beta, \tag{3}$$

where P_t symbolizes the nominal transmission power (dBM), K_e stands for a constant depending on the system, η signifies the path loss coefficient, d_1 is a reference distance for the antenna in far field, α denotes the fast fading effect, and β represents the random attenuation sparked by shadowing. The RSS is analyzed by adjusting d_0 which is the real transmitter-receiver distance.

The anchor nodes are nodes whose precise locations are recognizable prior to localization process. These can also be referred to as known nodes. When A represents the number of anchor nodes, a group of all WSN nodes with known locations is represented as KN. Thus, a known node position K_x is denoted by (p_{kx}, q_{kx}). Furthermore, unknown nodes are nodes whose location is calculated using a particular localization algorithm. The group of unrecognized WSN nodes is represented using UN:

$$\left.\begin{aligned} KN &= K_x | x = 1, 2, \ldots, A \\ UN &= U_x | x = 1, 2, \ldots, B - A \\ RN &= E_x | x = 1, 2, \ldots, C \end{aligned}\right\}. \tag{4}$$

Consider $B - A$ as the quantity of unrecognized nodes. In real-time request, the actual positions of U_x defined by (p_{ux}, q_{ux}) are undiscoverable. Let us assume the radius of the communication range to be R. Given that two energized sensor nodes are represented by p_x and p_y, if p_x is placed in the broadcasting range of p_y, p_x is directly considered a neighbor of p_y. As long as all the energized WSN nodes are endowed with very similar range of transmission, p_y is correspondingly a neighbor of p_x.

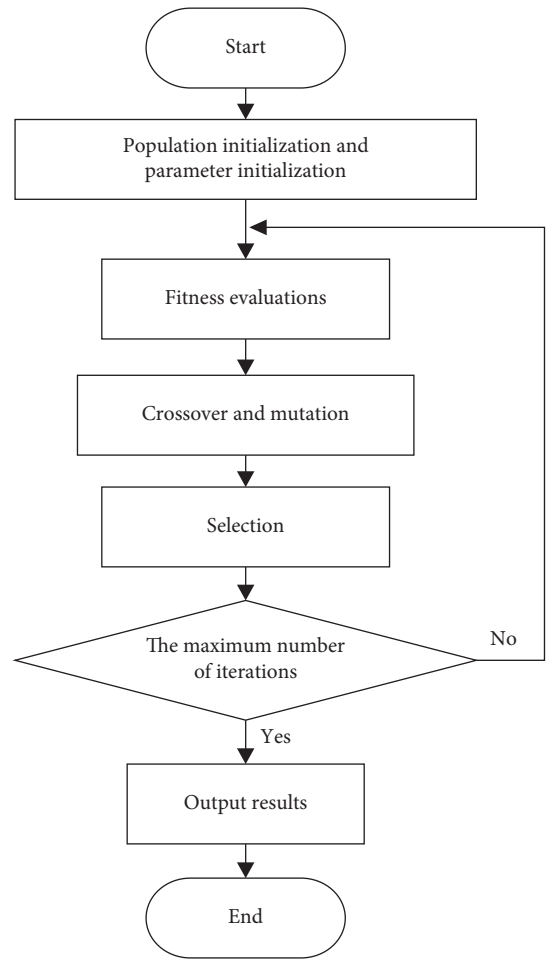


Figure 7: Flowchart for the genetic algorithm.

The likelihood of locating a node without a specific location is given as an estimated point because other unknown nodes are also being located during the positioning. The estimated position, actual position, and reference position are shown in Figure 8. The estimated position of U_x is represented by (p^0_{ux}, q^0_{ux}). Apparently, the rationale behind localization is to make $(p^0_{ux}, q^0_{ux}) = (p_{ux}, q_{ux})$ for U_x. The reference nodes comprise localized unknown nodes and known sensor nodes in the course of trying to locate an energized sensor node. The array of reference nodes is symbolized by RN, where $B \geq C \geq A$. The reference node point E_x with an exact position is defined by (p_{ex}, q_{ex}). In fact, if E_x is anchoring K_y, then $(p_{ex}, q_{ex}) = (p_{ky}, q_{ky})$. However, if E_x contains localized unknown nodes U_k, at the moment $(p_{ex}, q_{ex}) = (p^0_{uk}, q^0_{uk})$.

The actual distance which is represented by d_{xy} is the distance covered concerning the actual points of U_x as well as E_x. The measurement distance d^1_{xy} is obtained by a certain measurement method based on the error z, and this error is determined by the random used measuring instrument. For convenience, follow-up studies generally use random value instead of this error. We presuppose that

(1) Select: primary population
(2) Estimate: the individual capabilities of respective participant of the population
(3) Repeat
    Choose optimally performed participant to replicate
  Using a genetic operation (mutation and crossover), breed another generation and produce offspring
  Estimate the discrete capabilities of the reproduced offspring
  Substitute underperformed section of the population using the reproduced offspring
(4) Until <criteria are met>.

Algorithm 1: Pseudocode for the GA.

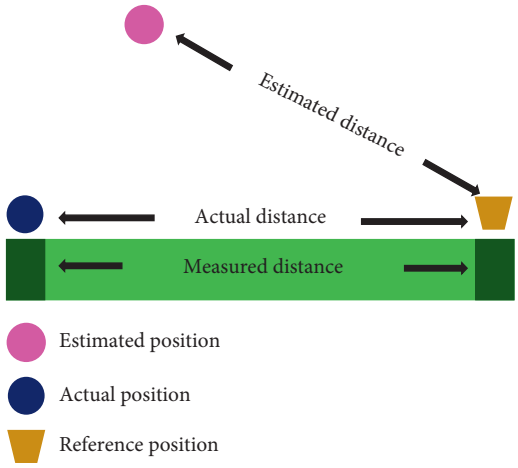


Figure 8: Distance between sensor nodes.

$d_{xy}^1 = d_{xy} + N(0, d_{xy}z)$, where $(N(0, d_{xy}z))$ is the Gaussian function with a mean value of 0 and a variance of $d_{xy}z$.

The distance covered by the estimated positions U_x and E_y is denoted by d_{xy}^0 (estimated distance). Suppose unknown node U_x has m neighbor reference nodes $E_1, E_2, \ldots, E_m$, where $y = 1, 2, \ldots, m$. We can derive the following equations to get (p_{ux}^0, q_{ux}^0):

$$d_{xy}^1 = \sqrt{\left(p - p_{ey}\right)^2 + \left(q - q_{ey}\right)^2}, \tag{5}$$

where (p, q) is an unknown dimension to resolve and p_{ey}, q_{ey} is the position of E_y. Due to the existence of the distance measurement error z and the estimated position (p_{ux}^0, q_{ux}^0) it is impossible to know the actual point (position) of E_y in a broad sense. Making use of the estimated position (p_{ux}^0, q_{ux}^0), the estimated distance d_{xy}^0 at that point is expressed as

$$d_{xy}^o = \sqrt{\left(p_{ux}^o - p_{ey}\right)^2 + \left(q_{ux}^o - q_{ey}\right)^2}. \tag{6}$$

Because of the uncertainty of d_{xy}, that is, the actual distance is different, and the goal of positioning is to achieve an infinitesimal distance from d_{xy}^1 to d_{xy}^0. Finally, we construct the location problem denoted U_x as

$$\sum_{y=1}^{n} w_y \left(d_{xy}^o - d_{xy}^1\right)^2, \quad \sum_{y=1}^{n} w_y \left(\sqrt{\left(p - p_{ey}\right)^2 + \left(q - q_{ey}\right)^2} - d_{xy}^1\right)^2, \tag{7}$$

where $w_y = (1/d_{xy}^1)\sum_{x=1}^{m}(1/d_{xy}^1)$, which gives better understanding pertaining to the reference point closer to U_x. In factual terms, the distance covering the communication range of an energized sensor is a halfway circle caused by multipath fading, asymmetrical message delivery, and redundant noise.

In finding the minimal localization error of unidentifiable (unknown) location points U_x, the change in estimated and actual location point should be always be considered which is found in the equation below.

$$LE_x = \frac{1}{R}\sqrt{\left(p_{ux}^o - p_{ux}\right)^2 + \left(q_{ux}^o - q_{ux}\right)^2}. \tag{8}$$

4.2. Clustering Model. During the course of node clustering, the most approximate energized sensors are observed to be in the same locality (cluster) which tries to save energy by reducing the transmission range and the closest point amidst the energized sensors. Figure 9 shows a setup of the suggested clustering scheme. The key concern is to discover the precise location which depends on several decisions on how to locate it. With the goal of finding the preferable location for a particular energized sensor, the distance of a sensor node is calculated using (5). Our new approach for well-organized clustering splits the entire WSN nodes using Euclidean distance connecting sensor nodes into numerous clusters. However, equal cluster size must be assured at some point in the clustering process. Taking one cluster into consideration, the sensor nodes are however placed in order to minimize (14), that is, the Euclidean distance between the location points and their immediate central point. Therefore, if the location point is initiated with a sensing range R in a deployment area consisting of energized sensors at the central point, then it is said to be covered. Consequently, the distance of a location point p_x and the central sensor node at a point q_c should be less than or equal to the distance between a location point p_x and any energized sensor node at point $q_y, \forall_y = 1, 2$ to R, and it is mathematically represented as $\mathrm{d}(p_x, q_c) \leq \mathrm{d}(p_x, q_y)$.

5. Proposed Fitness Approach for ECGAL

In this section, we derive the fitness function for the proposed energy-efficient clustering and localization using a genetic algorithm.

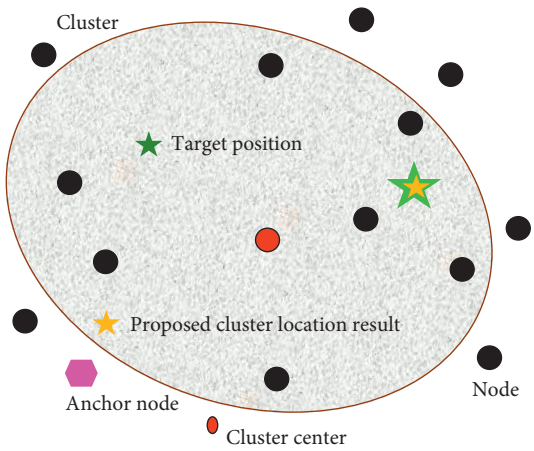


Figure 9: Expected WSN cluster structure.

5.1. Energy Efficiency. The ability for a chromosome to withstand all conditions helps it in lowering the energy exhausted and to maximize lifespan of the network system. The channel description for free-space and multipath fading used still considered the sum of distance connecting the receiver and the transmitter. Suppose the upper threshold value d_o is greater than the node distance pairs d, then the energy amplification consumption assumes a free-space model, but if d is greater than or equal to d_o, a multipath decay model is implemented. Therefore, the amount of energy needed by the radio to convey a h-bit message over a distance d is given in (9). The radio also uses up energy to accept a h-bit message given in (10). E_{elec} depends on factors for instance modulation, filtering, digital coding, and combining the dispersion of signals, but the amount of energy to amplify the system, $\varepsilon_{fs}d^2$ or $\varepsilon_{mp}d^4$, relies on the receiving structure per the distance travelled and the suitable error per bit. E_{elec} is defined as the electronic energy required by the electronic circuit and $d_0 = \sqrt{\varepsilon_{fs} \backslash \varepsilon_{mp}}$. ε_{fs} and ε_{mp} are the amplifier energies in free space and multipath, respectively. We denote E_i to be the residual energy once a message is communicated through h^{th}-bits at a distance d from the receiver in (11). E is the node's recent energy. Then, $E_1 = E_T(h,d) + E_R(h)$ is the energy needed to send a message plus the energy consumed while receiving a message.

$$E_T(h,d) = \begin{cases} h \times E_{\text{elec}} + h \times \varepsilon_{fs} d^2, & d < d_o, \\ h \times E_{\text{elec}} + h \times \varepsilon_{mp} d^4, & d \geq d_o, \end{cases} \tag{9}$$

$$E_R(h) = h \times E_{\text{elec}}, \tag{10}$$

$$E_i = E - E_1. \tag{11}$$

5.2. Distance Estimation. The total distance covered starts with an energized node point to another sensor point which is assumed to be the distance between two neighboring sensor nodes. It is expressed as $\text{dist}(p_a, q_b)$. However, the distance from a locational node to the central cluster position should be insignificant in order to get closer to our target compared to the distance from the cluster center to another node. The latter is represented by $\text{dist}(p_a, q_c)$ and the former is denoted as $\text{dist}(p_c, q_b)$. This should be ensured orderly so that the wastage of energy of each node in a large network is minimal. This boosts the cluster strength and reduces the lack of sensor node involvement. For all node points $p_x \varepsilon N$, where N is the set of all nodes, we compute the sum of the distance D_i with all its neighboring points q_y. However, these energetic neighboring points could be the position of a node without location which could be activated using the position of a known node. This distance is given in (14).

$$D_G^o = \sum_{p_a \in N}^{N} \text{dist}(p_a, q_b), \tag{12}$$

$$D_G^1 = \sum_{q_b \in N}^{N} \text{dist}(p_a, q_c) + \text{dist}(p_c, q_b), \tag{13}$$

$$D_i = \sum_{p_y \in N} \text{dist}\left(p_x, q_y\right). \tag{14}$$

5.3. Coverage Connection. Every WSN can be considered as a connected undirected figure denoted by $G = (V, E)$, where V consists of vertices comprising of $\{v_1, v_2, \ldots, v_u\}$ which denotes the energized node point found in the WSN along with E which is the edge set $\{e_1, e_2, \ldots, e_f\}$ representing the distance between the energized sensor nodes. This approach considers the weighted values depending on energy efficiency, distance estimation, and coverage connection which are represented on the edges. Furthermore, every edge in the network possesses a finite real number which is represented as w_i. Let the sensing range of a node be denoted by S_r. Let $c = c_1, c_2, \ldots, c_m$ be the connectivity variables associated with the energized sensor nodes p_x and p_y. However, C_y is the area covered by y^{th} cluster center node, N is the sum of all recognized energized sensor nodes, and C is the WSN area.

$$C_L = \begin{cases} 1, & \text{if } \left\| p_x - p_y \right\| \leq S_r, \\ 0, & \text{otherwise}, \end{cases}$$
$$C_i = \bigcup_{y=1}^{N} \in C_y \frac{C_y}{C}. \tag{15}$$

The final fitness function that demands to be computed in minimization is given below, and it constitutes the previous fitness minor objectives:

$$F_i = w_1 E_i + w_2 D_i + w_3 C_i, \tag{16}$$

where $w = \{w_1, w_2, \ldots, w_f\}$ is the distance associated with the edges. We define w_1, w_2, w_3 to be the weight coefficients attached to the fitness function in order to quantify its contribution to each of the other subfunctions, and it is expressed by $\sum_{i=1}^{3} w_i \geq 0, w_i \in (0,1)$.

6. Performance Evaluation

The performance of our approach is evaluated in this section. The device used for the evaluation was Intel(R) Core (TM) i5-3317U CPU PC with 6144 RAM, which was accomplished using MATLAB 2014a. It is compared with existing approaches such as DV-HOP (distance vector-hop), CENTA (centroid algorithm), EDV-HOP (evolutionary distance vector-hop), and CGAL (clustering in genetic algorithm localization). The network scenario is considered realistic in nature with 200 energized sensor nodes randomly deployed with 25% anchor nodes, several unknown nodes, and reference nodes. The experimental parameters deployed in this work are further presented in Tables 1 and 2.

Figure 10 provides a solid evidence to conclude that the new approach performs impressively compared to other location-based algorithms in error location. Almost all the applied approaches work robustly in the same configuration. ECGAL drops gently because of the additional anchor nodes in the network which provided more reference points for the target nodes. However, the network is boosted when there are sufficient anchor nodes because the distance joining the unknown nodes and the anchor nodes gets smaller. In our simulation output, CGAL, EDV-HOP, and CENTA showed less localization errors as well.

In Figure 11, it is assumed that as the transmission range expands, the number of seconds of continuous simulation operation increases. Simultaneously, the transmission range starts from 5 m and steadily increases by 5 m so as to evaluate the performance of our approaches. The ability to locate a node improves and could be achieved when the radius of transmission inclines which reduces the error in localization. Finally, when the transmission range increases, ECGAL obtains better results in terms of location accuracy.

The experimental results in Figure 12 show the task of location error computed against varying node numbers. On top of it all, as the number of the energized sensor nodes increases, the localization error for all the algorithms decreases slowly. Among all our localization approaches, ECGAL shows fewer points for its localization error. As the number of nodes reaches 200, more reference points are found which help to localize the node with less error. However, as there is increase in energized nodes, there is also slack in lifetime contributing factor for CENTA and DV-HOP.

Figure 13 shows the localization error against the number of clusters considering different algorithms. The clustering technique proposed in this paper improves the energy efficiency in the network. With the increase of cluster number, the localization error decreased. ECGAL and CGAL dropped slowly because when the number of clusters is high, fewer nodes will be found in their clusters, which makes it easier to locate an unknown node, thereby reducing the localization error drastically. However, CENTA is seen to perform almost similar to EDV-HOP because of its special clustering abilities. The energy depletion level of a network enhances if there are some reasonable number of clusters. In DV-HOP, the transmission scale has enough energized nodes, which indicates that more of these nodes are found in each cluster.

Table 1: Parameters used in sensor field and GA.

Simulation parameters	Value
Total number of nodes	200
Deployment field area	$200 * 200\,m^2$
Communication range R	40 m
Number of anchor nodes	50
Number of unknown nodes	100
Maximum iterations	300
Number of clusters	5
Population size	50
Length of chromosome	5
Number of generations	150
Mutation rate	0.5
Crossover % rate	0.8

Table 2: Parameters used in the energy model.

Parameters	Values
Initial energy	2 J
Distance (d_0)	87 m
Packet size	200 bits
Energy for transmitting (E_T)	50 nJ/bit
Energy for receiving (E_R)	50 nJ/bit
Energy for data aggregated (E_D)	50 nJ/bit/signal
Energy consumption of power amp in free space (ε_{fs})	$10\,pJ/bit/m^2$
Energy consumption of power amp in multipath fading (ε_{mp})	$0.0013\,pJ/bit/m^4$

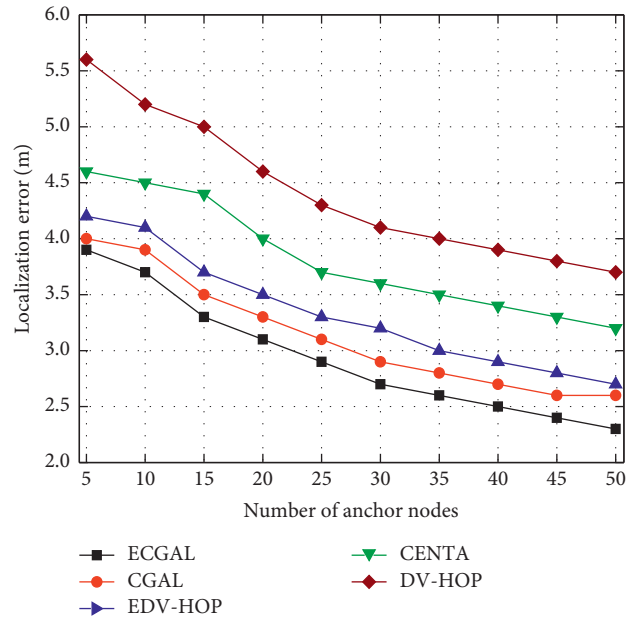


Figure 10: Localization error against the number of anchor nodes.

Figure 14 depicts the residual energy against the number of iterations. The energy savings are meaningful in ECGAL compared to CGAL, EDV-HOP, CENTA, and DV-HOP. All the approaches dropped more and more until 80 iterations where they start to drop significantly. The amount of residual

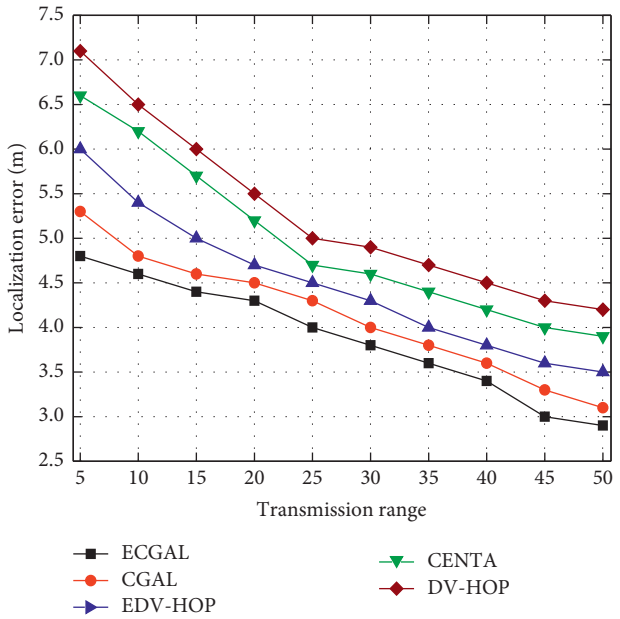


Figure 11: Localization error against transmission range.

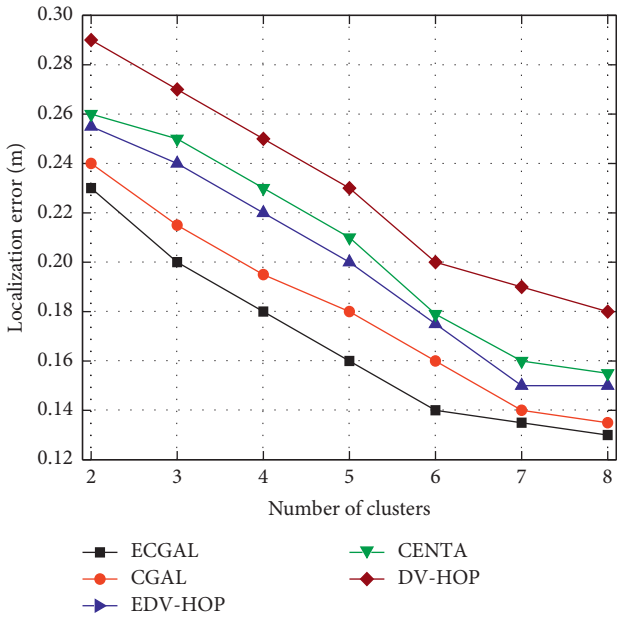


Figure 13: Localization error against the number of clusters.

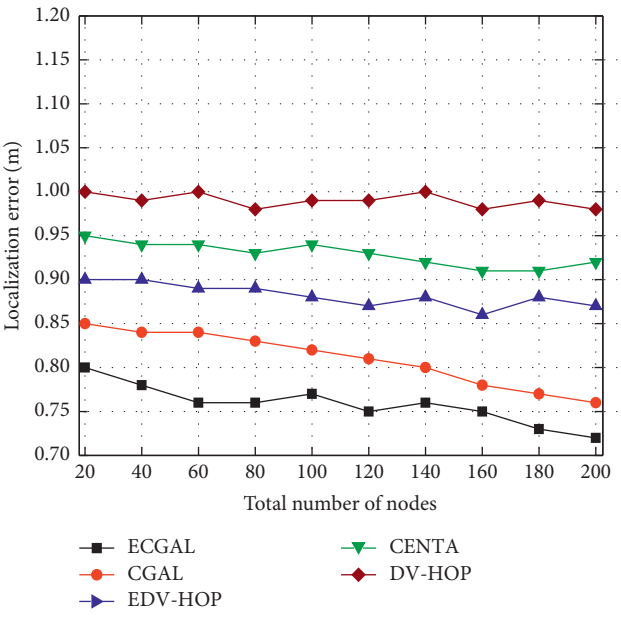


Figure 12: Localization error against the total number of nodes.

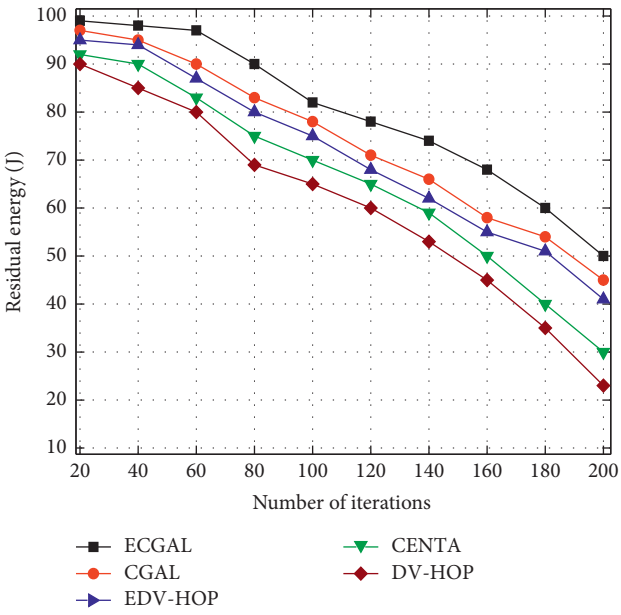


Figure 14: Residual energy against the number of iterations.

energy found in ECGAL surpasses CGAL because the energy left behind after 120 iterations is about 70 J. This could due to the optimal election of cluster centers and the equidistance between the intracluster and intercluster. At the initial position, the performances of CGAL and EVD-HOP are the same. According to our graph, the energy remaining for EDV-HOP and CENTA after 160 iterations was 55 J and 50 J, respectively. Finally, the decline in residual energy affects the life expectancy of the network which in turn increases the number of exchanged control packets (overhead).

In Figure 15, the time taken to process the algorithm is computed for 200 iterations over seconds. The number of messages generated to messages sent to the final point is described as the total success rate of packets delivered. Finally, ECGAL performs better in terms of the convergences rate which is best compared to CGAL, EDV-HOP, CENTA, and DV-HOP. It is clear that the ECGAL proves its success in transporting about 90% information to its final destination. With an increase in the number of iterations, CGAL and EDV-HOP showed better execution compared to CENTA and DV-HOP. The performance output of CENTA and DV-HOP is almost similar.

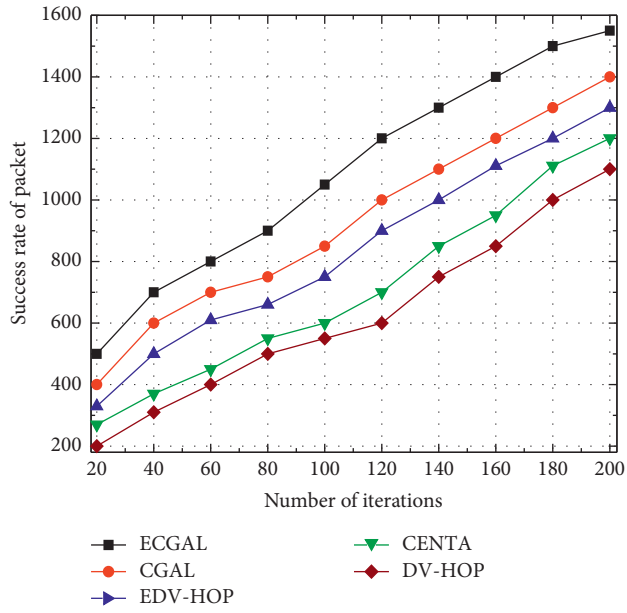


Figure 15: The success rate of packets against the number of iterations.

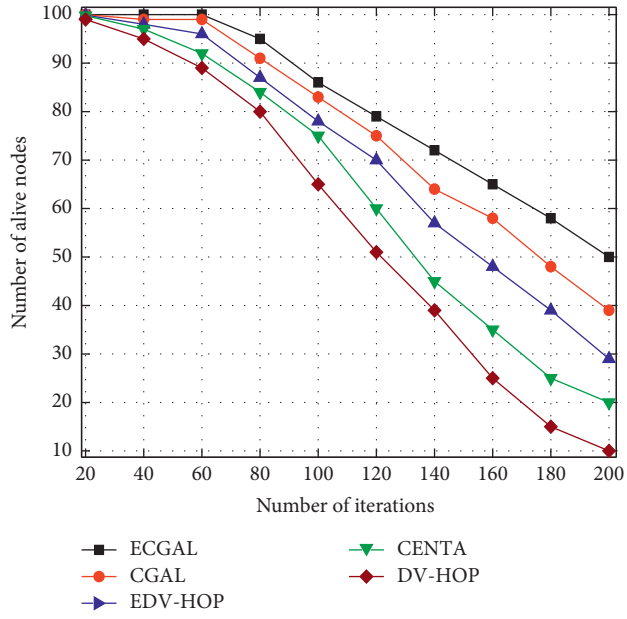


Figure 16: The number of alive nodes against the number of nodes.

Figure 16 shows the lifetime of the network. CGAL and the proposed technique radically improved the life of the network compared to EDV-HOP, CENTA, and DV-HOP techniques. The live nodes are evaluated by increasing the number of iterations to 200. The energy level of the networks' energized sensor nodes drains after several iterations. When the number of iterations reaches 180, compared with CGAL's 50 active nodes, ECGAL has 60 active nodes. At the same time, EDV-HOP and CENTA have only about 40 and 25 active nodes, respectively. The lines of the active nodes in the figure indicate that our method has a longer life span compared with other methods. On account of the developed

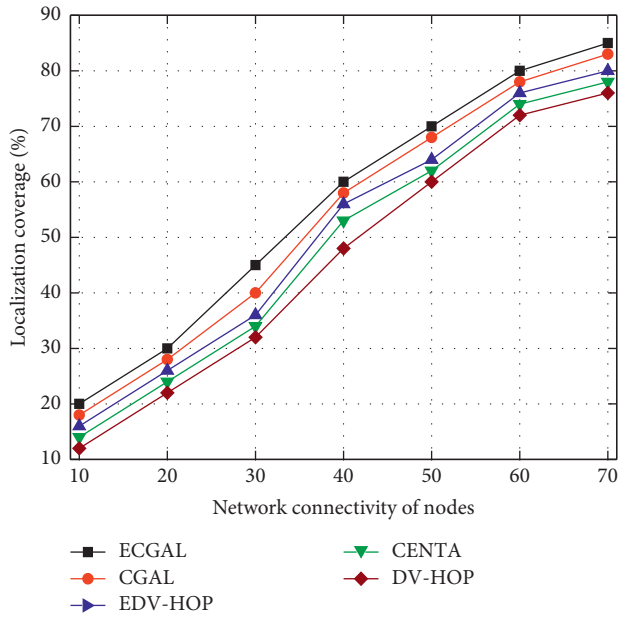


Figure 17: Localization coverage against network connectivity of nodes.

energy-efficient clustering localization approach based on genetic algorithm, ECGAL showed better output than that of the CGAL, EDV-HOP, CENTA, and DV-HOP algorithms.

In Figure 17, the influence of connected nodes is studied by analyzing the coverage against the network node correlation staring from 10 through 70. In our experiment, 200 energized sensors occupy 200 by 200 m^2 deployment area. With an increase in connected nodes, the localization coverage for ECGAL, CGAL, EDV-HOP, CENTA, and DV-HOP also increases. When the readings on horizontal axis reach 50 connected nodes, the network coverage still increases for ECGAL and CGAL. When the coverage location indicator accumulates 70 bars, the strength of the node to node relationship still goes higher for EDV-HOP, CENTA, and DV-HOP. The density of neighboring energized sensors warrants cost-effective and well-built connection between known and unknown node points. In general, our improved positioning method is superior to other existing methods in terms of accuracy.

7. Conclusion

The proposed ECGAL approach for ultimate location problem shows significant results after employing, and it proved that the energy-efficient clustering based on genetic algorithm localization approximates the node that demands to be identified and later assures a minimal location error when matched with DV-HOP, CENTA, EDV-HOP, and CGAL. ECGAL is better due to its efficient energy clustering strategy. In fact, our improved approach for better localization reconstructs our solution to quickly detect the location of the unidentifiable sensor node. However, nodes with known location point are randomly dispersed in an exact WSN because of the randomly deployed energized node point. For that reason, anchor nodes assist in locating

wherever unknown sensors are even though the converse is true, and thus excessive neighboring known points cause more nodes to be unlocalized. In conclusion, we can acknowledge that the proposed ECGAL performs effectively when studied with other approaches in relation to true position point and minimal error in terms of location.

Data Availability

The data used to support the findings of this study can be accessed from the following link: https://www.researchgate.net/publication/349836503_ECGAL.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Key R&D Program of China (no. 2018YFC0407101) and the Fundamental Research Funds for the Central Universities (no. 2019B22314).

References

[1] E. Fernandes, A. Rahmati, K. Eykholt, and A. Prakash, "Internet of things security research: a rehash of old ideas or new intellectual challenges?" *IEEE Security & Privacy*, vol. 15, no. 4, pp. 79–84, 2017.

[2] R. Kaur and S. Arora, "Nature inspired range based wireless sensor node localization algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 7–17, 2017.

[3] W. Lan, W. Zhang, and J. Luo, "Design and implementation of adaptive intelligent trilateral localization algorithm," *Chinese Journal of Sensors and Actuators*, vol. 30, no. 7, pp. 1089–1094, 2017.

[4] Y. Liu and J. Chen, "A *K*-means based firefly algorithm for localization in sensor networks," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 34, no. 4, pp. 364–379, 2018.

[5] P. Singh, A. Khosla, A. Kumar, and M. Khosla, "Optimized localization of target nodes using single mobile anchor node in wireless sensor network," *AEU - International Journal of Electronics and Communications*, vol. 91, pp. 55–65, 2018.

[6] S. K. Rout, A. K. Rath, P. K. Mohapatra, P. K. Jena, and A. Swain, "A fuzzy optimization technique for energy efficient node localization in wireless sensor network using dynamic trilateration method," *Advances in Intelligent Systems and Computing in Progress in Computing, Analytics, and Networking*, vol. 1, pp. 325–338, 2018.

[7] A. Azadeh, S. Goldansaz, and A. Zahedi-Anaraki, "Solving and optimizing a bi-objective open shop scheduling problem by a modified genetic algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 85, no. 5-8, pp. 1603–1613, 2016.

[8] S. Anthrayose and A. Payal, "Comparative analysis of approximate point in triangulation (APIT) and DV-HOP algorithms for solving localization problem in wireless sensor networks," in *Proceedings of 7th IEEE International Advanced Computing Conference*, pp. 372–378, Hyderabad, India, January 2017.

[9] T. Najeh, H. Sassi, and N. Liouane, "A novel range free localization algorithm in wireless sensor networks based on connectivity and genetic algorithms," *International Journal of Wireless Information Networks*, vol. 25, no. 1, pp. 88–97, 2018.

[10] S. P. Singh and S. C. Sharma, "A PSO based improved localization algorithm for wireless sensor network," *Wireless Personal Communications*, vol. 98, no. 1, pp. 487–503, 2018.

[11] S. H. Sackey, J. A. Ansere, J. H. Anajemba, M. Kamal, and C. Iwendi, "Energy efficient clustering based routing technique in WSN using brain storm optimization," in *Proceedings of 2019 15th International Conference on Emerging Technologies (ICET)*, pp. 1–6, Peshawar, Pakistan, May 2019.

[12] S. Sreenivasamurthy and K. Obraczka, "Clustering for load balancing and energy efficient in IoT applications," in *Proceedings of 2018 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pp. 319–332, Atlanta, GA, USA, August 2018.

[13] J. Wang, Y. Gao, K. Wang, A. K. Sangaiah, and S. J. Lim, "An affinity propagation-based self-adaptive clustering method for wireless sensor networks," *Sensors*, vol. 19, no. 11, p. 2579, 2019.

[14] S. H. Sackey, J. Chen, J. A. Ansere, G. K. Gapko, and M. Kamal, "A bio-inspired technique based on knowledge discovery for routing in IoT networks," in *Proceedings of 2020 IEEE 23rd International Multitopic Conference*, pp. 1–6, Bahawalpur, Pakistan, November 2020.

[15] Z. Yang, C. Liu, and L. Jin, "A clustering-based algorithm for device-free localization in IoT," in *Proceedings of 2018 IEEE 4th International Conference on Computer and Communications*, pp. 769–773, Philadelphia, PA, USA, December 2018.

[16] P. T. Daely and D.-S. Kim, "Bio-inspired cooperative localization in industrial wireless sensor network," in *Proceedings of 15th IEEE International Workshop on Factory Communication Systems*, Sundsvall, Sweden, May 2019.

[17] H. M. Kanoosh, E. H. Houssein, and M. M. Selim, "Salp swarm algorithm for node localization in wireless sensor networks," *Journal of Computer Networks and Communications*, vol. 2019, Article ID 1028723, 12 pages, 2019.

[18] F. Zhang, H.-F. Xue, Y.-H. Zhang, and F. You, "A new localization and tracking algorithm for wireless sensor networks based on the internet of things," *Sensors and Transducers*, vol. 154, no. 7, pp. 56–61, 2013.

[19] H. Musafer, R. Abdulhameed, E. Abdelfattah, and K. Elleithy, "A dynamic clustering algorithm for object tracking and localization in WSN," in *Proceedings of 27th International Conference on Computer Applications in Industry and Engineering, at New Orleans*, New Orleans, LA, USA, September 2014.

[20] C. Alippi, M. Bocca, G. Boracchi, N. Patwari, and M. Roveri, "RTI goes wild: RTI goes wild: radio tomographic imaging for outdoor people detection and localization," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2585–2598, 2016.

[21] T. Liu, X. Luo, and Z. Liang, "Enhanced sparse representation-based device-free localization with radio tomography networks," *Journal of Sensor and Actuator Networks*, vol. 7, no. 1, p. 7, 2018.

[22] C. Xu, B. Firner, Y. Zhang, and R. E. Howard, "The case for efficient and robust RF-based device-free localization," *IEEE Transactions on Mobile Computing*, vol. 15, no. 9, pp. 2362–2375, 2016.

[23] L. Lin and L. Donghui, "An energy-balanced routing protocol for a wireless sensor network," *Journal of Sensors*, vol. 2018, Article ID 8505616, 12 pages, 2018.

[24] J. Wang, D. Fang, Z. Yang et al., "E-HIPA: an energy-efficient framework for high-precision multi-target-adaptive device-free localization," *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 716–729, 2016.

[25] L. Zhang and T. N. Wong, "An object-coding genetic algorithm for integrated process planning and scheduling," *European Journal of Operational Research*, vol. 244, no. 2, pp. 434–444, 2015.

[26] S. H. Sackey, J. Chen, A. J. Henry, and X. Zhang, "A clustering approach based on genetic algorithm for wireless sensor network localization," in *Proceedings of 2019 15th International Conference on Computational Intelligence and Security (CIS)*, pp. 54–58, Macao, Macao, May 2019.

[27] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: a deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, 2017.

[28] S. Sivakumar, "Error minimization in localization of wireless sensor networks using fish Swarm optimization algorithm," *International Journal of Computer Applications*, vol. 159, no. 7, pp. 39–45, 2017.

[29] L. Song, L. Zhao, and J. Ye, "DV-hop node location algorithm based on GSO in wireless sensor networks," *Journal of Sensor*, vol. 2019, Article ID 2986954, 9 pages, 2019.

[30] G. Kumar and M. K. Rai, "An energy-efficient and optimized load-balanced localization method using CDS with the one-hop neighborhood and genetic algorithm in WSNs," *Journal of Network and Computer Applications*, vol. 78, no. 73–82, 2017.

[31] J. Wang, Y. Gao, X. Yin, F. Li, and H.-J. Kim, "An enhanced PEGASIS algorithm with mobile sink support for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 9472075, 9 pages, 2018.

[32] G. Sharma and A. Kumar, "Modified energy-efficient range-free localization using teaching-learning-based optimization for wireless sensor networks," *IETE Journal of Research*, vol. 64, no. 1, pp. 124–138, 2018.

[33] J. H. Anajemba, T. Yue, C. Iwendi, M. Alenezi, and M. Mittal, "Optimal cooperative offloading scheme for energy efficient multi-access edge computation," *IEEE Access*, vol. 8, pp. 53931–53941, 2020.

Hindawi
Complexity
Volume 2021, Article ID 6698231, 9 pages
https://doi.org/10.1155/2021/6698231

Research Article

AIBPO: Combine the Intrinsic Reward and Auxiliary Task for 3D Strategy Game

Huale Li, Rui Cao, Xuan Wang, Xiaohan Hou, Tao Qian, Fengwei Jia, Jiajia Zhang, and Shuhan Qi

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

Correspondence should be addressed to Jiajia Zhang; zhangjiajia@hit.edu.cn and Shuhan Qi; shuhanqi@cs.hitsz.edu.cn

Received 15 October 2020; Accepted 2 July 2021; Published 14 July 2021

Academic Editor: Abd E.I.-Baset Hassanien

In recent years, deep reinforcement learning (DRL) achieves great success in many fields, especially in the field of games, such as AlphaGo, AlphaZero, and AlphaStar. However, due to the reward sparsity problem, the traditional DRL-based method shows limited performance in 3D games, which contain much higher dimension of state space. To solve this problem, in this paper, we propose an intrinsic-based policy optimization (IBPO) algorithm for reward sparsity. In the IBPO, a novel intrinsic reward is integrated into the value network, which provides an additional reward in the environment with sparse reward, so as to accelerate the training. Besides, to deal with the problem of value estimation bias, we further design three types of auxiliary tasks, which can evaluate the state value and the action more accurately in 3D scenes. Finally, a framework of auxiliary intrinsic-based policy optimization (AIBPO) is proposed, which improves the performance of the IBPO. The experimental results show that the method is able to deal with the reward sparsity problem effectively. Therefore, the proposed method may be applied to real-world scenarios, such as 3-dimensional navigation and automatic driving, which can improve the sample utilization to reduce the cost of interactive sample collected by the real equipment.

1. Introduction

Machine game has always been one of the most active domains of artificial intelligence. According to whether the game states are fully observable, the machine game can be divided into two categories: perfect information game (PIG) and imperfect information game (IIG). The PIG refers to the game that participants can observe all game states, such as Go and Chess. On the contrary, the participants always hold private information in the IIG. Thus, the game state cannot be fully observed for the player, such as hole cards in Poker games. Deep reinforcement learning (DRL) has achieved great success in the field of the PIG recently [1–5]. However, due to the hidden information, the DRL-based methods have not shown satisfactory results in the IIG. The IIG is still a significant challenge in the field of machine games.

In recent years, many new testing platforms have emerged to verify methods of the IIG [6, 7]. In these platforms, the strategy game includes all the base elements of the IIG, such as StarCraft and VizDoom. Furthermore, the strategy game contains rich game scenarios and a considerable game state space [8–10]. Thus, the strategy game becomes a kind of ideal platform to verify relevant methods of the IIG. In view of the great success achieved by the DRL in the PIG, researchers have investigated intensively on solving strategy games through the DRL methods in the IIG. In 2019, DeepMind's AlphaStar, based on the DRL, became the first agent that defeated a professional human player in the StarCraft game, which has been viewed as a milestone in the advancement of artificial intelligence [11]. In addition, DeepMind designed an agent that can achieve human-level performance in a 3D multiplayer first-person video game, Quake III Arena in Capture the Flag mode, which only used pixels and game scores as the input [9].

Although the DRL-based methods have made certain achievements in the research of solving the strategy games, there are still many problems to be solved. The DRL methods highly rely on the reward to update the model during the

training process. Reasonable and instant rewards not only make the training process converge quickly but also make the learned model more robust. In the game scenario where the reward is sparse, the speed of the convergence will be slow down seriously and even more, resulted in a divergent model finally. In the 3D strategy games with a higher dimension of state space and more complex game scenarios, the reward sparsity problem will become more severe. Although there is a clear reward (e.g., the score of the game) in such game scenarios, most of the final goals need to be achieved by defining a series of subgoals in the early stage. Most of subgoals may not get rewards in time unless they are finished.

The reward sparsity problem makes the training process be much more difficult and slower because it cannot get a timely and effective reward. To solve this problem, many solutions such as reward reshaping, hierarchical DRL, and curriculum learning are proposed in recent years [12–18]. Among these methods, the intrinsic reward mechanism plays an important role. The design of the corresponding internal reward mechanism helps the agent update its policy according to the internal reward in the absence of environmental reward information. The definition of this RL model which includes intrinsic incentive reward is intrinsic incentive reinforcement learning (IIML). The general model of the IIML is shown in Figure 1. In the IIML, the external environment is the scene of the RL agent. The internal reward generation model is a fictitious environment. The agent can get the reward from external environment and intrinsic reward generation model, respectively. In this way, the agent gets more reward information than usual, which can accelerate the convergence of the agent.

However, in the process of practical application, these methods require carefully designed rewards for related problems or additional training data, which greatly limits the applicability of related methods. Therefore, these methods can not deal with the problem of sparse reward very well. To deal with the problem of reward sparsity, this paper improves the traditional strategy optimization DRL methods by introducing intrinsic reward mechanism. We proposed an intrinsic-based policy optimization (IBPO) algorithm, which improves the exploration performance of the agent in the 3D scene with the sparse reward. In this algorithm, the intrinsic reward is used to combine with the traditional strategy optimization. Moreover, by designing the various auxiliary tasks, we further proposed the auxiliary intrinsic-based policy optimization (AIBPO), which further improves the IBPO. The experimental results tested on the VizDoom show that our method has a better performance than the previous methods. We summarize our contributions as follows:

(i) We proposed a method, intrinsic-based policy optimization (IBPO) algorithm, to solve the reward sparsity problem in the 3D games with imperfect information. The IBPO effectively improves the exploration performance of the agent by combining the intrinsic reward and the traditional strategy optimization method.

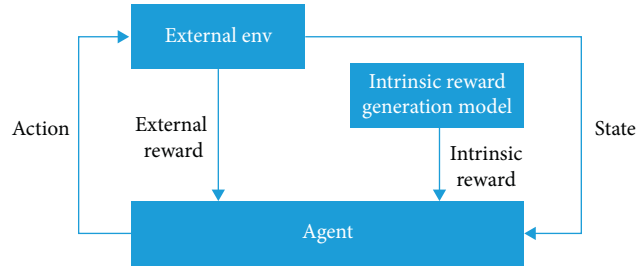


Figure 1: The model of the IIML (the agent takes the action according to the policy to obtain the external reward information fed back by the external environment; meanwhile, internal reward information will be generated in the internal environment no matter whether the current state agent gets external reward or not).

(ii) We presented the auxiliary intrinsic-based policy optimization (AIBPO), which is based on the IBPO, by integrating three kinds of auxiliary tasks: reward prediction auxiliary task, action value auxiliary task, and state value auxiliary task.

(iii) Extensive experimental results show that the performance of our method IBPO is significantly improved on VizDoom, compared with the previous methods. Moreover, the AIBPO further improves the performance of the IBPO.

The paper is organized as follows. The research method and its details are discussed in Section 2. Then, Section 3 shows the experimental results including performance comparsion with the state-of-the-art methods and ablation studies. Finally, the conclusion of this paper is given in Section 4.

2. Research Method

In this section, we seek a method of training the agent that can address the reward sparsity in the 3D game with complex environment. Intrinsic-based policy optimization (IBPO) is proposed by introducing an intrinsic reward mechanism. Moreover, combining with auxiliary tasks (reward prediction auxiliary task, action value auxiliary task, and state value auxiliary task), auxiliary intrinsic-based policy optimization (AIBPO) is presented based on the IBPO.

2.1. Intrinsic-Based Policy Optimization (IBPO). Due to the sparse reward problem in 3D strategy games, it is very difficult for the agent to carry out policy iteration and update. To overcome this problem, in this paper, we introduce the concept of intrinsic rewards, which can provide auxiliary reward information for the agent to update its policy. That is to say, here we employ an IIML (as shown in Figure 1) model. In this model, there are not only external rewards from the environment but also intrinsic rewards that are generated by a designed intrinsic rewards generation module. In addition, to adopt the intrinsic reward into the traditional policy optimization framework, we propose the intrinsic-based policy optimization (IBPO) algorithm, in

which the intrinsic reward generation module can be integrated with the policy optimization framework seamlessly.

2.1.1. The Generation of Intrinsic Reward. We describe how the intrinsic reward is designed. As mentioned above, the IIML contains the external reward and the intrinsic reward. The external reward is feedback from the environment, which cannot be obtained sometimes. When the agent sinks into a situation of lacking external reward, the intrinsic reward plays a key role in helping the agent to continue to update its policy. Especially, in the 3D game, if an external reward is sparse, the agent will easily fall into the problem of lingering in a local region. Therefore, to avoid the local region problem, the agent should increase its curiosity to explore unseen scenes. Based on this consideration, this paper constructs the intrinsic reward by estimating the novelty degree of the state. The novelty of a state means that the agent has not met this state before. A certain degree given to the new state can be viewed as the intrinsic reward, and the agent has a chance to overcome the local region problem.

Based on the above assumption, here we design the intrinsic reward generation module as follows. The intrinsic reward generation module is composed of a dual network structure as shown in Figure 2(a). In this structure, there is a target mapping network and a prediction network. Here, we set the target mapping network as a fixed network and the prediction network as a trainable network. Then, the intrinsic reward value is adopted as the similarity between the output vector of the target mapping network and prediction network. The input of these two networks is the current state of the game. That is to say, if the game state is novel, then the intrinsic reward will be a large value and vice versa. The main reason for this is to make the agent tend to explore unseen states in the environment. In the initial training phase, the agent has a smaller range of motion and there are more unfamiliar states in the game. In this case, the output vector of the two networks is less similar; thus, the calculated intrinsic reward value is larger. When the action policy is updated, the agent takes intrinsic reward as the main source of the reward.

Here, we describe the detail of constructing the target mapping network and the prediction network. A three-layer convolutional neural network is used to extract features from the input state, and finally a vector represented with a fixed dimension is output. The same network structure is adopted to reduce the error effect of different network structures on computing the similarity of the vector. The loss function L_{IR} of the intrinsic reward generation module is defined as follows:

$$L_{\mathrm{IR}} = \frac{1}{n}\sum_{i=1}^{n}(P-T)^2 + \lambda_{\mathrm{IR}}\sum_{j=1}^{k}\left\|\theta_j\right\|^2, \tag{1}$$

where P and T are the output vector of the prediction vector and the target vector, respectively, θ_j is the regularization term, λ_{IR} is the penalty factor of regular term, and n and j are the time steps.

2.1.2. Integrate the Intrinsic Reward into Policy Optimization. To train the intrinsic reward prediction network, adequate samples of the agent are necessary. Meanwhile, the agent's action in the environment is always driven by the policy optimization algorithm. In this way, it is a key to combine the intrinsic reward generation module with the policy gradient or policy optimization algorithm.

For the traditional policy optimization algorithm, the update of the policy mainly depends on the external rewards generated via the interaction with the environment. Therefore, to deal with the external reward and the internal reward is the key point of adapting the internal reward mechanism to the policy optimization algorithm. In this paper, we adopt a combination of long-term intrinsic reward and episodic external reward, as shown in Figure 2(b). The long-term reward can give the agent a goal. Meanwhile, it may lead to the reward sparse problem, which makes the strategy unable to converge or converge slow down. Therefore, the internal reward is introduced through the continuous incentive agent to explore, which will continue to support the agent to achieve its goal.

2.2. Auxiliary Intrinsic-Based Policy Optimization (AIBPO). The agent in 3D games typically requires a phased and long-term sequence to make decisions. The decision needs to be highly relevant to the current state, and the agent's action can make the state changeable. However, the state value of the DRL methods is estimated by the neural network, and the estimation is biased. From this perspective, we try to evaluate the state value and the action more accurately in 3D scenes. In this work, we designed three auxiliary tasks (reward prediction auxiliary task, action value auxiliary task, and state value auxiliary task) based on a multitask learning mechanism to perceive the information such as the reward and the connection between adjacent states, which can assist the agent in learning policy in 3D scenes.

The AIBPO with auxiliary tasks based on the IBPO is presented. Three kinds of auxiliary learning tasks are provided for the IBPO. The task can enhance agent's perception of the environmental reward to help the agent make decisions. The experience replay of the DRL is used, and the interactive sample is applied in the IBPO to train auxiliary tasks. The agent's decisions in 3D scenes are determined by learning and updating the parameters of the policy network in the DRL. The proposed auxiliary tasks can provide extra decision-support information for the agent. That is why the different auxiliary tasks can optimize the primary policy and auxiliary policy more effectively and robustly in 3D scenes.

The auxiliary task of training requires sampling the data that are related to the corresponding state, reward, and action. The IBPO agent stores these data in the experience replay memory during training process. In this paper, three kinds of auxiliary tasks (reward prediction auxiliary task, action value auxiliary task, and state value auxiliary task) are adopted by the AIBPO to assist decision-making. The auxiliary intrinsic-based policy optimization (AIBPO) can be defined as follows:

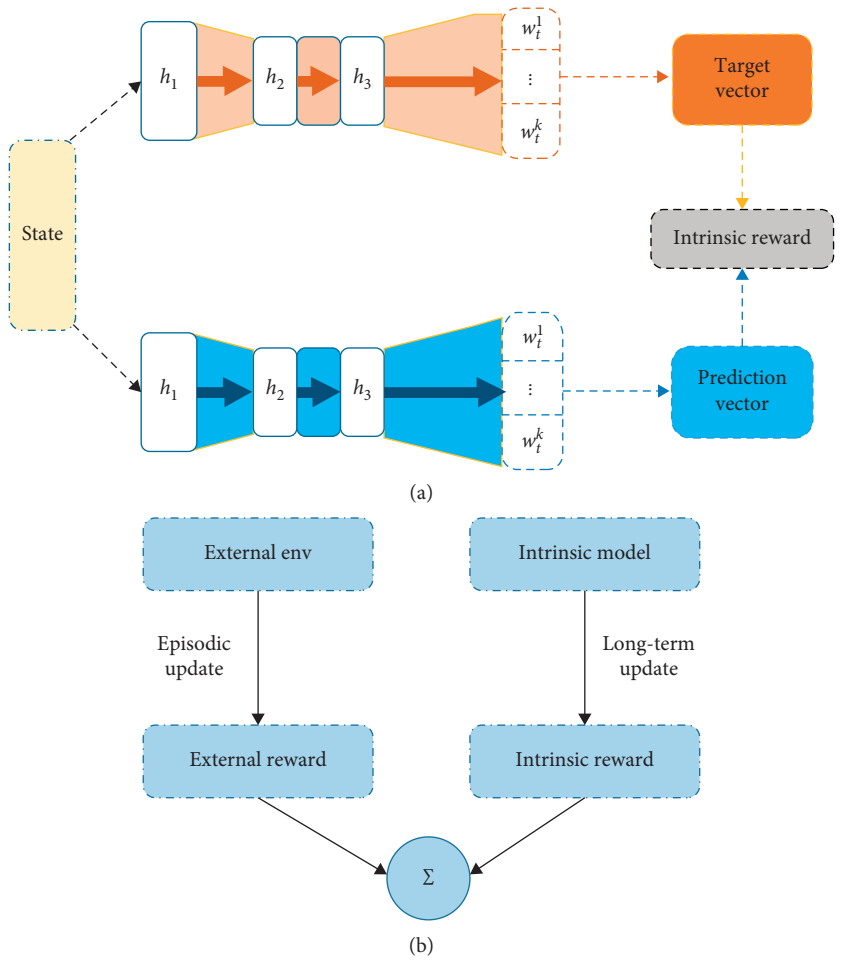


Figure 2: The structure of the IBPO (in subfigure (a), there are two channels: one is to output the target vector and the other is to output the prediction vector, which forms the intrinsic reward through the combination of them; subfigure (b) is the reward integration of the two rewards): (a) intrinsic reward generation module; (b) differentiated reward integration.

$$\mathscr{L}_{\text{AIBPO}} = \mathscr{L}_{\text{IBPO}} + \lambda_P \mathscr{L}_P + \lambda_V \mathscr{L}_V + \lambda_C \mathscr{L}_C, \tag{2}$$

where λ_P is the parameter of the reward prediction task, λ_V is the parameter of the state value task, λ_C is the parameter of the action value task, $\mathscr{L}_P$ is the loss function of the reward prediction task, $\mathscr{L}_V$ is the loss function of the state value task, $\mathscr{L}_C$ is the loss function of the action value task, and $\mathscr{L}_{\text{IBPO}}$ is the loss function of the intrinsic reward policy optimization algorithm.

Equation (2) gives the overall framework of the AIBPO. The AIBPO can learn useful information provided by different auxiliary tasks during training. This information is either related to the policy update or related to the scene perception, which improves the policy optimization in 3D scenes from different perspectives. Furthermore, the weight parameters in the auxiliary task loss function are set to determine the impact degree of the auxiliary task on the primary task. The integral model structure of the AIBPO is shown in Figure 3. The AIBPO is conducted by the whole process. The agent based on the IBPO interacts with the environment to generate more data, which are saved in the experience pool. The auxiliary tasks can be trained through sampling the experience pool.

Our goal is to add specific auxiliary tasks and finally pick up the best auxiliary tasks. To that effect, we prove that the optimal effect can be achieved by integrating the following three auxiliary tasks. The three auxiliary tasks are described in the following, respectively.

2.2.1. Reward Prediction Auxiliary Task. In the reward prediction auxiliary task, three consecutive frames are sampled as the network input. Through the convolutional layer and fully connected layer, the network outputs a classification category of rewards obtained by the agent, including positive rewards, negative rewards, and zero rewards. The label for the classification task is the one-hot encoding corresponding to the reward sampled from the experience replay memory at the next timestep. Because the multiclass cross-entropy loss function is used for classification tasks, here the loss function $\mathscr{L}_P$ in the reward prediction network can be defined as follows:

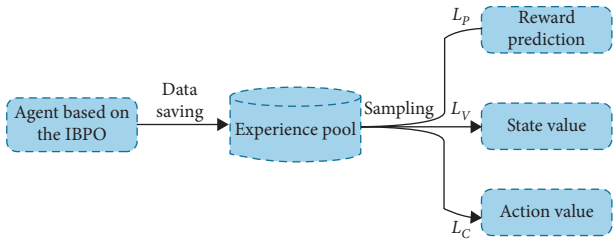


Figure 3: The model structure of the AIBPO.

$$\mathscr{L}_P = -\sum_{i=1}^{n} y_i \log\left(y_{i'}\right), \tag{3}$$

where y_i is the category of the network output and $y_{i'}$ is the reward value at the next moment.

2.2.2. State Value Auxiliary Task. The advantage estimation calculated by the method can be more accurate if the auxiliary policy can make the DRL methods tend to produce more accurate state value estimation. The auxiliary policy can make the training process of the agent more stable and the learning of the primary policy more efficient. Therefore, a state value auxiliary task whose regression prediction label is the state value corresponding to the image at the next timestep is presented. According to the mean square error loss function used for regression tasks, the loss function $\mathscr{L}_V$ in the state value network can be defined as follows:

$$\mathscr{L}_V = \frac{1}{n}\sum_{i=1}^{n}\left(v_i - v_{i'}\right)^2 + \lambda_V \sum_{j=1}^{k}\left\|\theta_j\right\|^2, \tag{4}$$

where θ_j is the regularization parameter, v_i is the state value of the network output, $v_{i'}$ is the target state value, and λ_V is the regularization penalty factor.

2.2.3. Action Value Auxiliary Task. The agent can be more prone to learn the primary policy that obtains the reward feedback efficiently if the auxiliary policy can make the selected action with greater value. Thus, an action value auxiliary task is presented here. In this task, the action value stored in the training process of the IBPO is sampled. The continuous frames and the temporal-difference action value can be taken as the data sample. According to the mean square error loss function used for regression tasks, the loss function $\mathscr{L}_C$ in the action value network can be defined as follows:t

$$\mathscr{L}_C = \frac{1}{n}\sum_{i=1}^{n}\left(Q_i - Q_{td}\right)^2 + \lambda_C \sum_{j=1}^{k}\left\|\omega_j\right\|^2, \tag{5}$$

where λ_C is the penalty factor of the regularization term, ω_j is the parameter of the regularization term, Q_i is the action value of the network output, and Q_{td} is the action value of the temporal difference.

3. Experimental Results

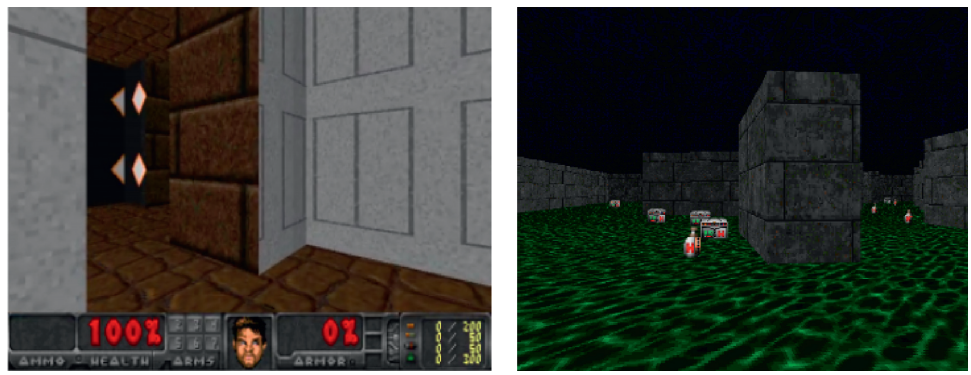
In this section, we first introduce the VizDoom. Secondly, implementation details are described. Finally, we conduct experiments to evaluate the performance of the IBPO and AIBPO. In addition, the ablation study of the AIBPO is also carried out.

3.1. Description of VizDoom. VizDoom is a first-person shooter game [8]. The agent's actions in the scene are similar to the real world, where they receive visual signals and then make decisions. As a mainstream research platform for the DRL methods, the VizDoom platform provides an interface to receive action input and reward signals and simulates the environment in a DRL model. Comprehensively, VizDoom is a platform capable of training agents to explore 3D environments. In this paper, experiments were conducted based on VizDoom's pathfinding and survival scenarios, which are described in Figure 4.

3.1.1. Pathfinding Scenario. Rewards are sparse in the pathfinding scenario, as shown in Figure 4(a). The agent can only get the reward at the target but not at any other location. The entire map is made up of several different opaque rooms with only one fixed target spot in a specific room. In this scenario, the agent can move freely but the starting position is far away from the target. The agent needs to pass through rooms with different contents to obtain rewards at the target spot.

3.1.2. Survival Scenario. There are mazes in the survival scenario, as shown in Figure 4(b), which block the agent from seeing a wide range of scenes. Moreover, the agent may continue to lose the life value during movement in this scenario. Therefore, it acquires the agent to explore and motion as efficiently as possible in the maze; otherwise, the training round would end when the life value turns 0. Besides, drug packages can restore life value appear randomly in the scenes. The agent needs to collect as many drug packages as possible to survive longer.

Different scenarios in the VizDoom platform have different environmental parameters and reward ranges. The primary evaluation indicators of the pathfinding scenario are the success rate of the agent's pathfinding and the number of required action steps. However, in a single training round, the agent cannot move on permanently. The maximum number of action steps is limited to 256 in the pathfinding scenarios. Once more than 256 steps, the current training episode ends immediately. The longer the agent spends in the pathfinding scenario, the weaker its exploration ability to be. Thus, the external environment will give the agent a penalty signal according to the time step. For the survival scenario, the primary evaluation indicators are the survival time step and the number of life value packages obtained. The agent's life value in survival scenarios decreases over time, and the decrease degree is presented by the platform.



(a) (b)

Figure 4: VizDoom's (a) pathfinding scenario and (b) survival scenario.

Table 1: The network structure.

NN1	Parameter	NN2	Parameter	NN3	Parameter
Conv	Conv (64, 3, 3)	Conv	Conv (128,3,3)	Conv	Conv (64, 3, 3)
Conv	Conv (32, 3, 3)	Conv	Conv (32,3,3)	Conv	Conv (32, 3, 3)
Active	ReLU	Active	BN	Fc	Linear (64, 3), reward prediction
Transform	Flatten	Active	ReLU	Fc	Linear (64, 3), reward prediction
Fc	Linear (288, 256)	Transform	Flatten	LSTM	Hidden state 256, state value
Fc	Linear (256, 256)	Fc	Linear (288, 256)	Fc	Linear (256, 1), state value
Actor	Linear (256, 4), a			LSTM	Hidden state 256, action value
Critic	Linear (256, 1), r			Fc	Linear (256, 1), action value
Critic	Linear (256, 1), r				

3.2. Implementation Details. The network structure of the adjusted policy optimization is shown in Table 1. The NN1 includes the actor network and the critic network [19], both of which share part of the hidden layer neurons. The critic network designed in this paper has two independent reward output heads to offer independent supervision information to the external reward and the intrinsic reward.

The intrinsic reward generation model makes up for the lack of the external reward when updating. The model includes a target mapping network and a prediction network, both of which have the same network structure. In Table 1, NN2 states the network structure of the intrinsic reward generation model.

The reward feature enhancement with auxiliary task learning enhances the agent's performance in 3D games by building their capacity on perception and state estimation. The samples for three proposed auxiliary tasks are obtained from the experience replay memory, which are stored when the agents interact with the environment. The network structures of the reward prediction, the state value, and the action value tasks are all composed of shallow convolutional neural networks and LSTM networks [20, 21]. In Table 1, NN3 shows the network structure of the auxiliary task.

In addition, the development platform of this research is Linux server with Ubuntu 16.04-5.4.0 system; the CPU is Intel (R) Xeon (R) Silver 4110 CPU @ 2.10 GHz, with 32 virtual cores; the GPU is NVIDIA Tesla P100, with 16 GB video memory; and the development language is based on Python 3.6.3.

3.3. Comparative Experiments

3.3.1. The Performance of the IBPO. The IBPO is tested in the pathfinding scenario. The agent starts to explore the entire scene from the starting position and cannot complete the task until it reaches the target position in the pathfinding scenario. This scenario is an extremely challenging training map in the VizDoom platform because there is only an external reward signal at the target. Another issue is that random agents have difficulty reaching the target position with limited action steps. Consequently, we test the IBPO algorithm in the pathfinding scenario.

In the pathfinding scenario, the evaluation indicators are the average reward value and average action steps. The average reward value is defined as the ratio of the number of the agent that reached the target position within the specified steps to the number of all trained agents currently, indicating the pathfinding success rate of the agents. The average action step is defined as the average number of action steps for 100 verifiable interactions between the environment and the agent after the algorithm converges, which indicates the stability of the agent's action policy.

We train different RL agents with IBPO, DRQN [22], and DFP [23], respectively, in the pathfinding scenario to prove the effectiveness of the IBPO by comparing their evaluation indicators. The experiments verify that the intrinsic reward can give agents assistance in updating policy

effectively without the environmental reward feedback. The DFP and the DRQN are the methods used by the second- and third-place agents in the 2017 VizDoom competition. In this competition, there is a map similar to the pathfinding scenario. The DRQN utilizes the game information provided by the simulator. The DFP owns supplementary processing modules for the state and the reward information. These two algorithms are dominant in the pathfinding scenario. Figure 5 illustrates the performance curves using the IBPO, DFP, and DRQN in the pathfinding scenario.

We can find that the IBPO reached an average reward value of 0.92, which outperforms the DRQN and the DFP, whose average reward values are 0.79 and 0.86, respectively. The closer the average reward value is to 1, the more effective the policy learned by the method is and the more likely the agent will reach the expected position. The main reason for this result is the deficiency of environmental rewards in the pathfinding scenario. The intrinsic rewards in the IBPO make up for the lack of positive reward values in experience replay memory, thus promoting the learning of exploration policy. The above comparative experiments demonstrate that the DRL agent trained by the IBPO has plentiful and efficient performance in exploring the 3D pathfinding scenarios.

Table 2 summarizes the average reward value and average action step in the pathfinding experiments. The first row corresponds to the reward in Figure 5, and the second row shows action steps required by different agents to reach the target position. We can find that the IBPO performs the best among the three methods with a minimum average action step of 61.8. It indicates the agent trained by the IBPO can find the path to the target faster with a more steady action policy.

In addition, in order to verify the effectiveness of the intrinsic reward in the IBPO, we also analyzed the trend of the intrinsic reward in the training process. As shown in Figure 6, the intrinsic reward value gradually decreased from 1 to about 0.1. In the whole training process, for the DRL agent, there are many novel states in the early stage of the training, which results in a larger intrinsic reward value. At this time, the intrinsic reward value is the main reward signal for the RL to update. As the training process goes on, the agent gradually learns the action policy to explore the environment and the internal reward value gradually decreases. At this time, the update of the RL algorithm is mainly external reward.

To sum up, the value generated by the intrinsic reward module provides auxiliary signals for the agent's policy update, which helps the agent to learn effective exploration policy in sparse reward scenarios. Through comparative experiments with the DRQN and the DFP, we conclude that the IBPO outperforms these two methods in the average reward value and average action step indicators.

3.3.2. The Performance of the AIBPO. The AIBPO is tested in the survival scenario. In VizDoom's survival scenario, the layout of each training round is different. The initial position, the number of drug packages, and the packages' locations are also random. Moreover, the maze wall limits the vision of the agent. Such random factors in the scenario would affect the agent's policy learning. The update will be unstable during training if the algorithm is unable to perceive the environmental information adequately.

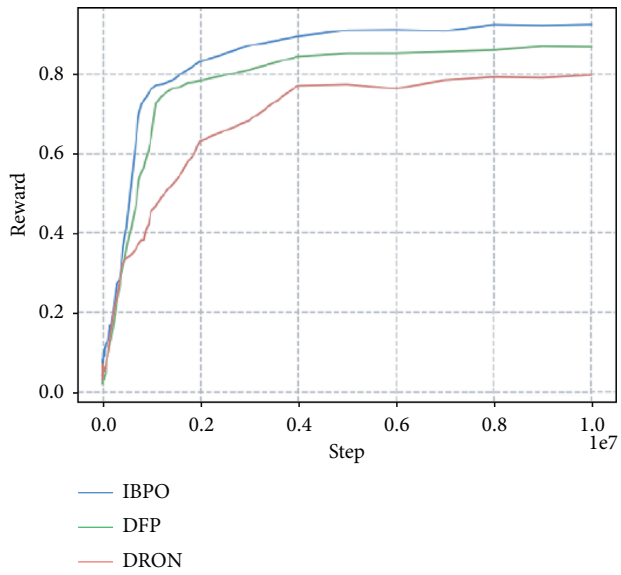


Figure 5: Experimental results of the IBPO in the pathfinding scenario (the y-axis represents the average reward value obtained by the DRL agents in the pathfinding scenario and the x-axis represents the timestep during training; the greater the reward, the better the method).

Table 2: Experimental results of the IBPO (the bold is the best).

Evaluation criteria	IBPO	DFP	DRQN
Average reward value	**0.92**	0.86	0.79
Average action steps	**61.8**	69.3	75.7

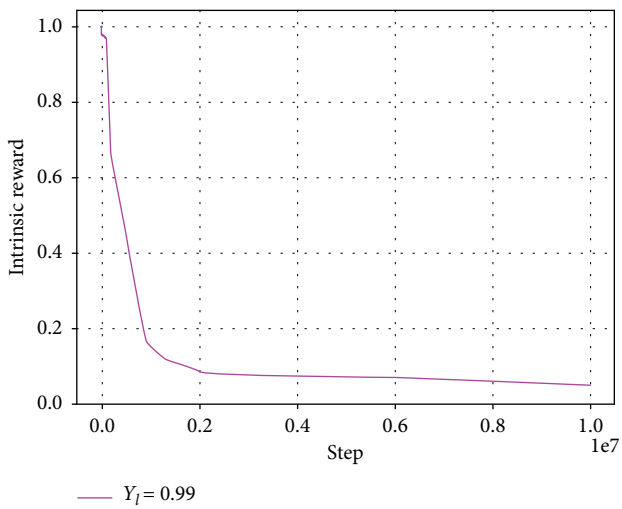


Figure 6: The trend of the intrinsic reward value (the y-axis represents the intrinsic reward value, and the x-axis represents the timestep during training).

Table 3: Experimental results of the AIBPO in the survival scenario.

Evaluation criteria	AIBPO	IBPO	DRQN	DFP	A3C	Rainbow
$\text{Surs}_{\text{step}}$	**293.7**	266.5	277.1	301.3	256.1	260.3
Pics_{hp}	**9.58**	7.83	7.41	8.81	6.85	7.32

Table 4: Ablation experiment results of auxiliary tasks.

Evaluation criteria	AIBPO	IBPO	IBPO + $\mathscr{L}_P$	IBPO + $\mathscr{L}_V$	IBPO + $\mathscr{L}_C$
$\text{Surs}_{\text{step}}$	**293.7**	266.5	273.4	268.7	282.5
Pics_{hp}	**9.58**	7.83	9.11	8.19	8.17

The AIBPO, DRQN [22], DFP [23], A3C [19], and Rainbow [24] are used to train the DRL agents for comparative experiments in the survival scenario according to the relevant evaluation indicators. The AIBPO is an improved version based on the IBPO with the auxiliary task learning mechanism. Its main body is still the IBPO; thus, an IBPO agent was also implemented in this experiment. $\text{Surs}_{\text{step}}$ and Pics_{hp} are the evaluation indicators that are equivalent to game scores in the DRL. The $\text{Surs}_{\text{step}}$ is defined as the average of the maximum action steps that the agent can survive for 100 verifiable interactions after the algorithm has converged. The Pics_{hp} is defined as the average number of drug packages picked up by the agent for 100 interactions after the algorithm has converged. Table 3 shows the experimental results in the survival scenario. The bigger $\text{Surs}_{\text{step}}$, the better Pics_{hp}.

We can find that the AIBPO is 9.58 on the Pics_{hp}, which is better than all other compared methods. They are 7.83, 7.41, 8.81, 6.85, and 7.32 for the IBPO, the DRQN, the DFP, the A3C, and the Rainbow on the Pics_{hp}, respectively. Meanwhile, the IBPO is also better than other methods except the DFP on the $\text{Surs}_{\text{step}}$. In addition, the IBPO does not perform better than the DFP and the DRQN on the $\text{Surs}_{\text{step}}$. The reason is that the IBPO principally aims at the sparse reward problem in 3D scenes, but there is sufficient reward information in survival scenarios. Fortunately, the maze features also require agents to make long-term plans based on the historical information. The sufficient reward information in this scenario can be used by auxiliary tasks of the AIBPO. Three types of auxiliary tasks play key roles in increasing the agent's reward perception and state estimation capacity. As a result, the AIBPO achieves better scores than the IBPO and the DRQN in this experiment. It is close to the DFP on the $\text{Surs}_{\text{step}}$ and superior to the DFP on the Pics_{hp}.

The proposed IBPO aims to solve the exploration problem in 3D scenes with sparse reward. The experimental result from the pathfinding scenario demonstrates that the IBPO has a certain efficiency improvement compared with the DFP in this type of scenes. The proposed AIBPO supplements the IBPO's shortcomings in perceiving environmental reward information by using an auxiliary task learning mechanism. The experimental result from survival scenarios shows that the auxiliary task learning mechanism as an auxiliary method has greatly improved the IBPO.

3.3.3. Ablation Experiment. The proposed AIBPO includes three types of auxiliary tasks: reward prediction task $\mathscr{L}_P$, state value task $\mathscr{L}_V$, and action value task $\mathscr{L}_C$. We attached the three tasks separately to the IBPO to compare their individual effects on the original policy optimization algorithm. The experimental results are shown in Table 4.

As we can see from Table 4, in survival scenarios, each of three auxiliary tasks has improved the benchmark algorithm IBPO to varying degrees, thus prolonging the survival time of the agent. The reward prediction task performed better on the Pics_{hp} than the other two tasks, while the action value task worked best on the $\text{Surs}_{\text{step}}$. Practically, they work better together than separately. The AIBPO that integrates all three tasks improved the most in the 3D circumstance.

3.4. Discussion. In many real-world scenarios, the agents require to make decisions in 3-dimensional space, for example, 3-dimensional navigation and automatic driving. Our work is able to be directly applied to these tasks, so our method shows very important application value. In addition, in some tasks based on reinforcement learning, reward sparsity is a common problem that limits the performance of the algorithm. For example, in the task of manipulator grasping, the manipulator can only get the reward after successfully grasping the target by completing a series of complex pose control. The failure of any step in the middle progress may lead to the failure to get the reward. Our approach provides additional rewards to the agent through intrinsic rewards and auxiliary tasks so as to alleviate the problem of reward sparsity effectively. Therefore, our method has a strong reference significance for these tasks in theory.

4. Conclusions

We have presented a novel approach, named IBPO, for the reward sparsity problem in 3D games. Unlike existing DRL-based methods, an agent of our approach can learn the intrinsic reward, which uses the differential fusion mechanism and the modified value network. Moreover, the AIBPO is proposed based on the IBPO by combining auxiliary tasks, which further improves the performance of the IBPO. The experimental results based on the VizDoom platform show the effectiveness of the proposed approach.

However, this method also has its limitations. First, it needs considerable expert knowledge in designing intrinsic rewards and auxiliary tasks, which limits its further application. Second, when used in more complex scenarios, computer vision, situation analysis, and other techniques are needed to make our method more robust.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by PINGAN-HITsz Intelligence Finance Research Center, Research and Development Plan of Key Fields in Guangdong Province, China (No. 2020B0101380001), National Natural Science Foundation of China (No. 61902093), and Natural Science Foundation of Guangdong (No. 2020A1515010652).

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Playing atari with deep reinforce-ment learning," 2013, https://arxiv.org/abs/1312.5602.

[2] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[3] D. Silver, A. Huang, C. J. Maddison et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[4] D. Silver, J. Schrittwieser, K. Simonyan et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[5] J. Schrittwieser, I. Antonoglou, T. Hubert et al., "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[6] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: an evaluation platform for general agents," *Journal of Artificial Intelligence Research*, vol. 47, no. 7, pp. 253–279, 2013.

[7] J. M. Font and T. Mahlmann, "Dota 2 bot competition," *IEEE Transactions on Games*, vol. 11, no. 3, pp. 285–289, 2018.

[8] M. Wydmuch, M. Kempka, and W. Jaskowski, "Vizdoom competitions: playing doom from pixels," *IEEE Transactions on Games*, vol. 11, no. 3, pp. 248–259, 2018.

[9] M. Jaderberg, W. M. Czarnecki, I. Dunning et al., "Human-level performance in 3D multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019.

[10] K. Shao, Y. Zhu, and D. Zhao, "Starcraft micromanagement with reinforcement learning and curriculum transfer learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 1, pp. 73–84, 2018.

[11] O. Vinyals, I. Babuschkin, W. M. Czarnecki et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[12] N. Hanaki, "Action learning versus strategy learning," *Complexity*, vol. 9, no. 5, pp. 41–50, 2010.

[13] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3249–3260, 2016.

[14] Z. Yang, K. Merrick, L. Jin, and H. A. Abbass, "Hierarchical deep reinforcement learning for continuous action control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5174–5184, 2018.

[15] Z. Guo, X. Wang, S. Qi, T. Qian, and J. Zhang, "Heuristic sensing: an uncertainty exploration method in imperfect information games," *Complexity*, vol. 2020, Article ID 8815770, 9 pages, 2020.

[16] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman, "Teacher–student curriculum learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3732–3740, 2019.

[17] J. Li, X. Shi, J. Li, X. Zhang, and J. Wang, "Random curiosity-driven exploration in deep reinforcement learning," *Neurocomputing*, vol. 418, pp. 139–147, 2020.

[18] N. Bougie and R. Ichise, "Skill-based curiosity for intrinsically motivated reinforcement learning," *Machine Learning*, vol. 109, no. 3, pp. 493–512, 2020.

[19] V. Mnih, A. P. Badia, M. Mirza et al., "Asynchronous methods for deep reinforcement learning," in *Proceedings of the International Conference on Machine Learning*, pp. 1928–1937, New York, NY, USA, June 2016.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[21] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] G. Lample and D. S. Chaplot, "Playing fps games with deep reinforcement learning," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 2140–2146, San Francisco, CA, USA, 2017.

[23] A. Dosovitskiy and V. Koltun, "Learning to act by predicting the future," in *Proceedings of the International Conference on Learning Representations*, pp. 637–645, Toulon, France, 2017.

[24] M. Hessel, J. Modayil, H. Van Hasselt et al., "Rainbow: combining improvements in deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1147–1153, New Orleans, LA, USA, 2018.

Hindawi
Complexity
Volume 2021, Article ID 5511866, 26 pages
https://doi.org/10.1155/2021/5511866

Review Article

Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics

Min-Hua Chao, Amy J. C. Trappey, and Chun-Ting Wu

Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan

Correspondence should be addressed to Min-Hua Chao; s106034803@m106.nthu.edu.tw

Received 19 February 2021; Revised 2 May 2021; Accepted 14 May 2021; Published 25 May 2021

Academic Editor: Abd E.I.-Baset Hassanien

Natural language processing (NLP) is a critical part of the digital transformation. NLP enables user-friendly interactions between machine and human by making computers understand human languages. Intelligent chatbot is an essential application of NLP to allow understanding of users' utterance and responding in understandable sentences for specific applications simulating human-to-human conversations and interactions for problem solving or Q&As. This research studies emerging technologies for NLP-enabled intelligent chatbot development using a systematic patent analytic approach. Some intelligent text-mining techniques are applied, including document term frequency analysis for key terminology extractions, clustering method for identifying the subdomains, and Latent Dirichlet Allocation for finding the key topics of patent set. This research utilizes the Derwent Innovation database as the main source for global intelligent chatbot patent retrievals.

1. Introduction

Despite the global impact of COVID-19, almost 80% of global artificial intelligence (AI) projects have maintained the same or even increasing the investments of R&D since the beginning of the pandemic. AI-based systems nowadays are widely adopted for decision makings, which have a profound impact on individuals and society. The so-called intelligent systems are mostly driven by machine learning (ML) or deep learning (DL) algorithms with their models being trained and tested by big data [1]. As an important application of AI technologies, smart chatbots (or called intelligent chatbots) help answer a large number of questions related to the pandemic [2]. Statistics provide reliable insights into trends in the intelligent chatbot development. Reported by Business Insider, the market size of chatbots is expected to grow from US$2.6 billion in 2021 to US$9.4 billion in 2024, with a compound annual growth rate (CAGR) of near 30% [3]. The study shows that more than 50% of customers, in various business sectors, expect businesses to be open 24/7. Chatbots, or virtual agents, enable company organizations to answer and take care simple questions and requested tasks from call centers, help desks, and service agents, and at the same time pass more complex issues to the real staffs and personnel, thereby controlling the human resource costs. Chatbots can save up to 30% of customer support costs with shortened response time and answering up to 80% of regular questions [4].

The applications of intelligent chatbots have increased rapidly in recent years. A lot of research delves into the details of AI and DL algorithms for chatbot solutions and applications in pursuits of high efficiency and intelligence. Even though the development of chatbot seems to be booming, thorough review of the life cycle of chatbot developments and key technologies are in great needs. Furthermore, with the popularity of the Internet and social platforms, a digitally transformed environment for the uses of smart chatbots (as human machine communication interfaces) has become largely popular. More and more applications offer "life" services by mounting voice-interactive assistants; that is, smart chatbots, which hold regular conversations and provide online services interactively with

users, are becoming a trend [5]. As a technology pioneer or market leader, based on the comprehensive review, one can identify innovative technologies or applications to maintain its lead. As a company that wants to follow the trend of digitization and entering into chatbot applications, betting resources on the most valuable development and finding the right breakthrough are the best strategy, knowing the state-of-the-art technologies and applications through the review article. To bridge the research gap, this research aims to use the intelligent ontology extraction and patent-mining methodology to review comprehensive chatbot-related patents and their innovative technologies and applications.

Chatbot is a computer program that allows computers to mimic human communications and conversations. At first, chatbot can only answer standard questions where questions and answers are known and saved in the system. With the technological advances, computers can gradually answer a freelanced question like human by passing a Turing Test, which is closer to a human intelligence [6]. With the rapid development of AI in recent years, intelligent chatbot has entered a new era and has been widely applied in many industries. For example, the voice customer query interfaces of large shopping malls, bank chatbots for monthly account balance queries, and even the well-known Siri reflect how chatbot technology gradually enters into people's daily lives through intelligent interfaces. NLP is becoming the norm for obtaining information, allowing companies to easily obtain key information from text documents, thereby enhancing operational efficiency or improving service levels. NLP also has many applications in other fields. Taking the medical industry as an example, NLP technology detects signs of cognitive impairment by analyzing the conversations between the elderly and patients with Alzheimer [7]. In the banking industry, optical character recognition (OCR) and NLP technologies are used to automatically capture key document text and perform document content reviews to speed up the lending process [8]. For the catering services, NLP is used to analyze customers' comments and emotions for improving services or performing precision marketing [9].

NLP-enabled chatbot is a complex system. Starting from the front-end user inputting utterance, the natural language understanding (NLU) module of chatbot judges the user's intent from the user's natural language expression. Next, the dialogue management module finds contents that can answer the user's request. In this process, different types of databases may be accessed for finding answers. Finally, the natural language generation (NLG) module converts the collected contents into human-readable expression as the response to the user [10]. NLP-enabled chatbot is also a smart system that integrates many AI technologies. The chatbot technology that uses AI to imitate human conversations has begun to mature and provides accurate solutions or answers to complex questions. Because natural language-enabled chatbots have the ability to map oral or written inputs to intent, they become popular in many applications, such as in manufacturing or service industry. Before there were chatbots, when employees wanted to obtain data from the company's information system, they needed to log into the system, select the corresponding function, find the corresponding file searching through complex file folders, and then finally access the needed information. With a chatbot, a single verbal request can complete the task. Among enterprise-level applications, there are few voice-enabled chatbots, but the demand for such functions is increasing. In addition, on the premise of satisfying basic service functions, soft functions are essential to the success of chatbots. Chatbots that incorporate features such as tone, emotion, and personality are desirable. Furthermore, smart chatbots tolerate human errors or allow fuzzy requests, still generate accurate answers, and are very attractive [11].

NLP technology is an important branch of AI. It studies the use of computer software, such as machine learning (ML), to intelligently process natural language. The basic NLP technology is mainly developed around seven levels of language, including phonemes (language pronunciation patterns), morphology (words, how do letters form words, the morphological changes of words), vocabulary (the relationship between words), syntax (how words form sentences), semantics (the corresponding meaning of language expression), pragmatics (semantic interpretation in different contexts), and chapter (how sentences are combined into paragraphs).

As AI drives the transformation of the digital economy, companies should also pay more attention to intellectual property (IP) innovation and management. Therefore, it is expected that the latest trend of chatbot development can be found from collective patent information. Through the patent layout (or landscape), important technology development trends can be evaluated, and the development direction of important international manufacturers can be found, and international technology benchmarks can be used as a reference for subsequent R&D investment decisions [12].

According to statistics from the World Intellectual Property Organization (WIPO), more than 80% of emerging technologies with commercial values are patented, which shows that the patent database consists of comprehensive domain knowledge. The purpose of the patent database is not only to provide a search for prior arts, but also to obtain a wealth of information for future R&D. For example, when key patents are found, the technology development trends can be extrapolated, the technical contents of domain patents can be analyzed, and the core countries, assignees, and inventors of the key technologies can be identified. By making good use of such patent information, companies can develop various business and management strategies [13].

In order to understand the latest emerging technologies of chatbots, this study takes "natural language-enabled chatbots" as the domain for relevant patent technology exploration. Thus, the overall chatbot technological development trends can be discovered and future research directions can be suggested.

Before investigating natural language-enabled chatbots, a well-constructed knowledge ontology is needed. Afterwards, the global patent management landscape map and technology function matrix are presented. After that, a discussion of the analytical results will be presented to show

the interesting technology trends we found and verified with the matching literature. In this study, some text-mining tools are used, such as clustering and topic modeling. Saura [14] summarized the types of 11 analysis methods of data sciences (DS) in digital marketing and provided good support for the patent-mining analysis method used in this study.

2. Literature Review

2.1. Patent Review Workflow. Past patent reviews are usually analyzed by experts. However, with the increasing number of patents and the development of information technology [15], most patent reviews are now performed through text-mining technology. Even with the assistance of text-mining technology, if there is no systematic patent review workflow, it is likely to cause the deviations from the subject of patent analysis. Abbas et al. [16] present an overview of the research workflow and tools on patent analysis. They divide the patent review workflow into three parts, including preprocessing, processing, and postprocessing. Retrieving patents and transformation into structured data are for preprocessing. Extraction of structures is for processing, including key term extraction and specific statistical data. Patent analysis approaches are for postprocessing that is classified into two categories, text-mining-based approach and visualization approach [17]. Kim and Bae [18] present a method for forecasting emerging technology of health care by patent analysis. They define the patent review workflow that can be divided into four stages, domain patent acquisition, technology clustering, technology defining, and evaluating patent clusters. They also mention that technology clustering results may vary depending on the analyst. In order to avoid a lack of objectivity, they focus on Cooperative Patent Classification (CPC) for forecasting emerging technology. In terms of the nonpatent literature, it is also an ample source of analyzing emerging technologies. Thilakaratne et al. [19] present a literature-based research workflow. They define the article retrieval process in detail for avoiding missing any related articles, including the retrieval rules and the standards of selection. In the article retrieval process, defining main research purposes, key words, and searching strategies are three parameters for determining the patent database. After constructing the patent database, they use systematic criteria to determine the literature is relevant or not. There have three stages for filtering the literature: the first stage is analyzing title and abstract, the second stage is analyzing introduction and conclusion, and the last stage is a complete reading and using a quality checklist. After that, visualization techniques are used to present their findings. In summary, the entire patent review workflow can be summarized into three main parts, patent search for determining the database, patent analysis for extracting key information, and result display for presenting the result in an easily understandable way.

Govindarajan et al. [20] proposed a systematic research flow for industrial immersive technology. Start by identifying the domain definition and confirming the scope of the research, and then after the main domain technical review, then keyword identification and ontology generation are carried out. The method cross-references a large number of technical articles and essential patents, ensuring a high coverage of technical information in specific fields. Finally, with text-mining technology, LDA topic modeling method, and TFM, a complete research flow structure is formed.

2.2. Patent Database. In a knowledge-based economy, the economic status of a country depends on the production, distribution, and use of knowledge and information. The latest trend of economic growth in various countries mainly depends on the individual's innovative technological knowledge, which is an important reason why intellectual property has attracted attention. Information related to intellectual assets, such as technical insight and legal status, cannot be obtained from any other literature search except for the patent database. Thus, the importance of the patent database can be revealed [21]. Krejcar et al. [22] compared several common large-scale patent databases, including AcclaimIP, Symphony Innovate, Inteum, IPzen, FoundationIP, Thomson IP Manager, and Derwent Innovation (DI), and pointed out the power of DI. The DI database uses the scientific literature, global patent data, and commercial data, so it can make more confident decisions in IP. Powerful analysis functions and simple workflow tools make DI be the best solution.

Derwent World Patents Index (DWPI) and the smart search function are two major features of DI. DWPI is a process of translation, rewriting of key abstracts, content debugging, and normalization of patent holders after experts have read the entire official patent disclosure materials, which is considered to be the essence of the patent content. The DWPI rewritten items include novelty, use, advantage, technical focus, detailed description, drawing description, activity, and mechanism. Every operation of DI simultaneously searches the official patent publications and DWPI patent value-added database to obtain more complete results. This is also the unique feature of DI. Smart search will analyze the word string semantically and automatically expand keywords, and then go through multiple steps of calculation, including weighting of classification numbers and weighting of citations, to find patents related to the input technical description. Grammar is not that important here, because smart search will remove conjunctions, prepositions, etc. in the description and only retain the technical keyword description. Therefore, whether the words used in the technical description are accurate or whether they are mixed with too many unnecessary technical conditions have more influence on the search results than the grammar. If the keywords left by smart search after analyzing the string are not as expected, or the results found by smart search for the first time do not meet the requirements, manually adjust, including adding new keywords in the search pane, or removing possible noise to let smart search recalculate new results. After several adjustments, the result of smart search will be closer to the demand. Smart search is an iterative process, the purpose is

to quickly find potential targets, and if you want to search all related patents without omission, it is suitable to use general patent search technique [23].

2.3. Ontology Construction. An ontology map for a specific domain connects the relevant subjects and key terms, provides a domain knowledge-rich structure that can be as the basis for analyzing technologies in depth. Weng et al. [24] presented a lexicon-based ontology construction method, which utilized term frequency and weighted factor to identify the relationship between key terms. If a term has significant weight, then it will be imported to the lexical database. The critical words for constructing the ontology are selected from the lexical database. Trappey et al. [25] proposed an information extraction approach and a knowledge-based ontology construction method for smart retailing technology mining, in which unsupervised ML methods are applied, including clustering and Latent Dirichlet Allocation (LDA), to construct a complete ontology by continuously refining. Tsatsou et al. [26] proposed an automatically constructing ontology method, which utilized the term frequency-inverse document frequency (TF-IDF) technique to determine key terms that may be branches or nodes of the ontology. Subhashini and Akilandeswari [27] mentioned that constructing an ontology is required to follow the six key steps, determining the scope of the ontology, capturing related data, encoding those useful data to machine-usable, integrating the results, evaluating the results, and documenting the ontology. In summary, constructing an ontology can mainly be divided into three parts, data source, determining the relationship between terms, and effectiveness evaluation.

2.4. Patent Mining. Patent documents contain important research results. However, they are lengthy and rich in technical terms, so analysis requires a lot of manpower, and there is an urgent need for automatic tools to assist patent engineers or decision makers in patent analysis. The importance of patent mining is thus seen. Patent-mining technology includes text segmentation, abstract extraction, feature selection, term association, cluster generation, topic identification, and information mapping [28]. In addition to the extensive use of LDA topic modeling methods in ontology construction, it is also very popular in patent mining.

In the patent analysis application of drones, through LDA, the three most active technology development themes such as communication technology, power supply, and navigation system are found [29]. Based on LDA, Korobkin et al. [30] proposed a new patent-mining method, which includes statistical and semantic analysis of patent documents, machine translation of patent applications, and calculation of semantic similarity between patents and applications. In the aspect of term association, Hu et al. [31] utilized a skip-gram-based model to extract key terms from patents and compared the proposed approach with the TF-IDF method. In terms of cluster generation, k-means is still powerful. Shanie et al. [32] used the k-means method to cluster patent documents related to green tea, in which the adaptive cluster number determination method is adopted based on silhouette score. Recently, ML methods for patent analysis have also begun to appear. Li et al. [33] proposed a DeepPatent that combines the convolutional neural network (CNN) model with the word embedding model for classifying patents. Lee and Hsiang [34] fine-tuned a bidirectional encoder representations from transformers (BERT) model to classify patents and compared the fine-tuned model with the previously mentioned model, DeepPatent, and the result shows that the precision is 9% higher. Jun [35] proposed a method for technical integration and analysis using boosting (an ML algorithm that can be used to reduce bias in supervised learning) and ensemble learning. This method uses regression trees, random forests, extreme gradient enhancement, and ensemble models After analyzing the integrated patent data, it can be extended to technology integration and analysis in more than three technical fields.

2.5. Technology Function Matrix (TFM). To further focus on the patent development context of a specific technical field and find a technical minefield or a technical blue ocean zone, it is necessary to analyze the technical location and function of each patent through a more detailed TFM, and further explore in-depth strategies, such as technological innovation or avoiding development conflicts [36]. In the patent analysis of cyber-physical systems (CPSs) and Industry 4.0, Trappey et al. [37] adopted domain ontology and International Patent Classification (IPC) as the basis of TFM. However, IPC and Cooperative Patent Classification (CPC) are general classifications. When exploring technology in a specific field, a large number of patent documents may have the same or similar classification codes, which makes the identification of technical classifications insufficient, and finally manual interpretation by professionals is still required. According to a survey of examiners at the European Patent Office (EPO), 84.7% of examiners believe that CPC is very important for patent searches. Although 70% of examiners believe that AI and ML technologies can provide valuable support in the future, about 45% of examiners still believe that patent searches fundamentally rely on human efforts. And 52% of examiners do not think that a fully automated patent search can be done before 2035 [38].

In the practice of the industry, most of the patents collected are read by the researcher one by one and classified according to the technical field and effect of their professional human judgment. The manual classification method consumes a lot of time, and it is difficult to obtain a comprehensive review through the interpretation of a large number of patent documents. Many recent studies have tried to find a more efficient way to construct TFM. Yang and Ren [39] proposed a semiautomatic TFM construction method by extracting technical words and computer-aided algorithms to reduce labor costs and time. Ki and Kim [40] proposed a programmatic automation method based on NLP technology to quickly construct an Information Relation Matrix (IRM), which describes relationships among technical information in the patent and is similar to TFM. Trappey et al. [41] used the resultant patent text and data

mining technology to create ontology-based TFM for patent analysis of additive manufacturing in the dental industry. The abundant literature shows that ontology, text mining, NLP, topic modeling, and TFM technology can be regarded as the main procedures for patent analysis today.

2.6. Comparison. Table 1 shows the comparison of the 17 related studies of technology, especially patent-mining techniques from its research purposes, tasks for preprocessing, processing, and postprocessing [16]. The second column of Table 1 lists the tasks in each part, and the third column lists the more specific methods used. Each successive column corresponds to each article, of which part, task, and method used are listed.

In the preprocessing part, the use of natural language processing for text preprocessing is mentioned in most articles, and the corresponding algorithms, tools, or kits are quite mature. Although some articles did not specifically mention this part, it is believed that this part, as a relatively mature part, should have been implemented. Two main tasks, key term extraction and patent management map, are included in the processing part. The TF-IDF method is widely used in the key term extraction task and can almost be regarded as a standard configuration. Skip-gram is an important method to study the contextual relationship, and it is often used in the research that uses the contextual relationship as the vectorization method. Patent management map, or patent map analysis, is a statistically-based data analysis method that has been widely used, with a database and business intelligence tools to visualize patent portfolios. Patent management map only involves data sorting and presentation, which does not conform to the current general definition of text-mining. Therefore, it is hardly mentioned in the research of patent analysis by text-mining in recent years. Among them, only the patent classification code will be referenced as a benchmark to verify whether the results of the text-mining-based approach are valid and consistent. The postprocessing part contains two parts: text-mining-based approach and visualized approach. The main methods of the former are clustering, topic modeling, and classification; the latter is mostly based on the expression of node-relation graph. Although TFM is less common, it is still one of the good visualization tools for exploring emerging technologies.

The main purpose of these studies is focused on classification, ontology construction, and finding emerging technologies. Classification is very basic, and the patent data itself already have classification codes, such as IPC or CPC. Researchers who use classification methods in post-processing parts have a clear aim at classification. Ontology construction aims to clarify the technical details and scope of a specific field, and clustering and topic modeling methods can achieve this goal well. Both classification and ontology construction only obtain and analyze existing data, but in order to explore emerging technologies, it is necessary to find rules or discover changes in trends from the data.

The framework proposed in this study completely includes the three parts of preprocessing, processing, and postprocessing. In addition, this research also performed patent management map analysis and compared the results with text-mining to explore emerging technologies and verify the ideas and conclusions put forward in this research.

3. Patent-Based Ontology Construction

Figure 1 illustrates the ontology construction process, including four levels and two aspects.

The four levels are patent retrieval, patent clustering and target domain selection, topic modeling, and keyword generation. The two aspects are research process and ontology construction. At level 1, some key terms about natural language-enabled chatbot are figured out, and the smart search on DI is used to do the patent retrieval. Then, the most related 50 patents are quickly glanced to check if they match the subject of this study. If not, the search query is adjusted and do the retrieval again until the records are much in line with the subject. At level 2, DWPI title, DWPI abstract, and independent claims are used to do the k-means clustering, and silhouette score is used to evaluate the propriate number of clusters. After clustering, normalized TF-IDF (NTF-IDF) is used to identify the key words and key phrases. Again, we will check if the key words match the subject. If not, go back to level 1 and adjust the search query. Repeat the process until ideal target domains are found. At level 3, topics for domain are found in 2 different ways. The LDA model is used in domain of NLP, model, and system, while manual induction is used in domain of applied scenarios. In order to discover deeper topics or concepts at this level, each domain resets patent search conditions for applying the LDA method. After each execution, it is determined whether the subject of each domain is clearly identified according to the results. If not, the patent search conditions must be adjusted again. The topics of each domain are determined in this iterative process. Finally, by sorting out the key words and key phrases from level 2 and level 3, the construction of level 4 can be completed.

3.1. Patent Retrieval. Smart search on DI provides a semantic search tool, which offers a quick path to capture related patents from simple search terms. The powerful algorithm behind replicates the strategies used by expert searchers to provide a manageable result set that matches users' intent. By using smart search, it is not necessary to list all probable related terms before searching. Instead, the records discovered are always related to the technology described by the input terms but may not be exactly contained. Smart search automatically sorts the result set according to the relevance score to show the content that best matches the search term.

In order to obtain a well-constructed ontology, the main purpose is to find as wide a range of technologies as possible from the field, and not to focus on specific technologies that will lead to a small number of emerging technologies that cannot be found. Smart search has the advantage of intelligence, but the limit of 1,000 records corresponds to about 450 to 550 DWPI families on average, which is not much in

Table 1: Comparison for studies related to patent and technology mining techniques.

Part	Task	Method	Proposed framework	[18]	[24]	[25]	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[37]	[39]	[40]	[41]
Preprocessing	Data processing	Text preprocessing	✓		✓	✓		✓	✓	✓	✓								✓	✓
Processing	Key term extraction	TF/TF-IDF	✓		✓	✓	✓		✓	✓	✓	✓	✓			✓	✓			✓
		Skip-gram	✓								✓	✓		✓						
	Patent management map	Yearly trend	✓	✓						✓										
		Assignee	✓							✓										
		CPC/IPC	✓	✓													✓		✓	
Postprocessing	Text-mining-based approach	Clustering	✓	✓		✓	✓					✓	✓							
		Topic modeling	✓		✓	✓			✓	✓	✓									
		Classification			✓				✓			✓		✓	✓	✓	✓		✓	
	Visualized approach	Semantic tree analysis									✓									
		Node-relation graph	✓		✓	✓	✓													
		TFM	✓														✓	✓	✓	✓
Purpose	Prior art patent search									✓										
	Classification							✓			✓	✓	✓	✓				✓		
	Ontology construction	✓		✓	✓	✓	✓													
	Finding emerging technologies	✓	✓						✓						✓	✓			✓	

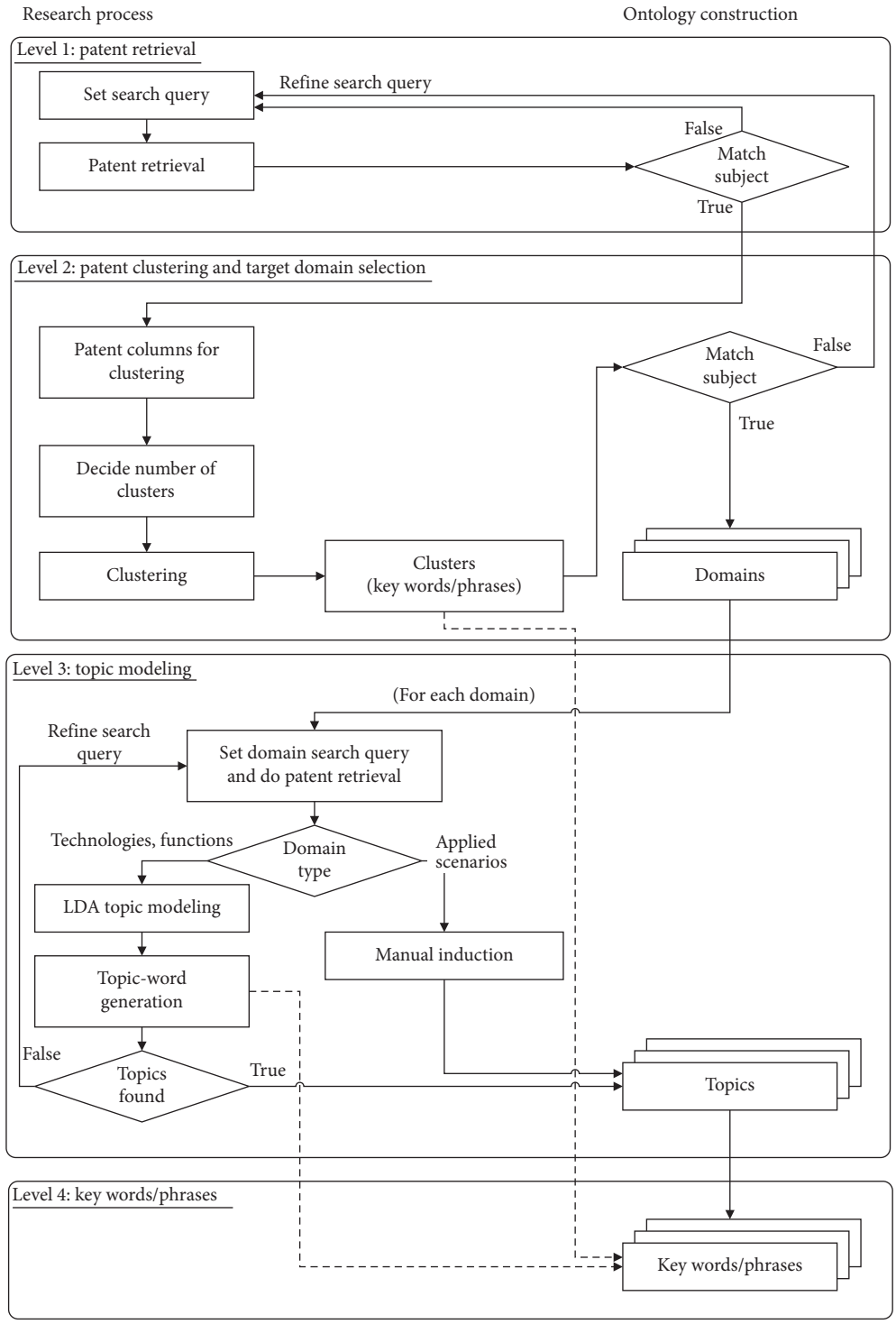


Figure 1: The ontology construction process flow.

terms of the number of patents related to NLP chatbot. The results of patent search will be used as the data source for clustering task at level 2. To use more patents for clustering, traditional patent search is also tried, which directly search patents form the original term user lists. Although by traditional search more patents can be found, if there are emerging technologies or applications that are not widely discussed or even undetected, they will not be found. After several rounds of trials this study finally selected 508 DWPI families detected by smart search as the results of level 1, and its search is shown in Table 2.

3.2. Patent Clustering and Target Domain Selection. At level 2, the patent obtained from the previous level is clustered and some target domains are discovered from the results.

Table 2: Search query for clustering.

Search type	DI query	Result
Smart search	SSTO = ("natural language processing" "natural language understanding" "NLP" "NLU" "chatbot" "VIRTUAL ASSISTANT" "INTELLIGENT ASSISTANT" "automated conversational interface")	508 DWPI families

The process begins at extracting the words in the patent document and using NTF-IDF to do vectorization, so that numeric vectors are obtained and can be applied to perform the k-means clustering. After that, the top words and n-gram top phrases of each cluster can be counted, from which target domains are selected.

3.2.1. Patent Columns for Clustering. This study chooses DWPI title, DWPI abstract, and independent claims as the source attributes for clustering. Patent documents may come from different countries, written in different languages, and cover a large number of attributes. Patent is to protect the inventor's smart finance or as a consideration for the enterprise's knowledge layout. Contrary to academic articles, patents are not written for users to understand easily, and some information may even be deliberately hidden in the title, which is not conducive to patent mining. The DWPI title and DWPI abstract, provided by the DI database, just solve the above problems. DI employs discipline-professional editors with scientific and engineering backgrounds to manually read all patents one by one and rewrite the title and abstract with easy-to-understand text, which are DWPI title and DWPI abstract, respectively. They remove the legal jargon, use American spelling, and intellectually choose drawing instead of just choosing the ones on the front page. In addition, many studies have shown that the value of patents is greatly affected by the number of independent claims, which are also included as the source attribute of the cluster.

3.2.2. Clustering. After retrieving and vectorizing patent documents, k-means can be performed to show the clustering distribution phenomenon in the vector space. The appropriate number of clusters can be obtained by calculating the silhouette score: the goal is to maximize the distance between clusters and minimize the distance within clusters. In this study, 13 clusters are clustered from 508 patents, and the top 10 words and 2-gram phrases in each cluster are extracted through NTF-IDF (see Table 3).

3.2.3. Domain Selection. The top 10 words and 2-gram phrases of 13 clusters, with a total of 260 terms, of which technical details are examined individually, are classified as 13 subdomains, which are combined to form the 4 domains, that is, NLP, model, system, and applied scenarios (see Table 4). The subdomains can display related topics and assist topic selection when performing topic modeling in level 3. "NLP" domain is distributed in clusters 3, 4, 5, 6, 8, 9, 11, 12, and 13; "model" domain is concentrated in cluster 6; "system" domain has clusters 7, 9, and 10, and "applied scenarios" domain is in clusters 1 and 2. Some clusters are related to multiple domains at the same time. Since the purpose of cluster analysis is to find out target domains, it is not so important whether each group must be clearly assigned to only one certain domain.

This research takes natural language-enabled chatbot as the subject. A large number of words related to NLP appear in large numbers in each cluster, which is not helpful to find out the domain, such as "NLP," "natural language," and "processing." In addition, many chatbot-related words are very versatile, which also increase the difficulty of domain exploration, such as "processor," "request," "input," and "module." The above vocabularies are skipped during the domain selection. One step in the preprocessing of patent documents before clustering is to vectorize the patent documents. Although those skipped terms in Table 4 could be set as stop words in the preprocessing stage, the reason for not skipping them is to avoid affecting the integrity of some phrases. Take "recognition" as an example. "Recognition" is also included in "intent recognition," "named entity recognition," "speech recognition," and "image recognition." While setting "recognition" as a stop word, the above related phrases will not be found. However, failing to remove "recognition" has caused it to appear repeatedly in each cluster and does not have domain recognition.

"NLP" domain contains cognition, named entity recognition (NER), linguistics (which include syntactic, semantic, and morphology), natural language understanding (NLU), response, and speech recognition. Nine clusters, cluster 3, 4, 5, 6, 8, 9, 11, 12, and 13, are distributed in NLP domain. For cluster 3, two subdomains, cognition and response, are involved. For cognition subdomain, representative patent US9361884B2 (assignee: Nuance Communication Inc.) proposed a human-machine dialogue system, incorporating with an NLU engine and a dialogue manager for providing NLP application to identify and resolve anaphora. For response subdomain, patent US10417266B2 (assignee: Apple Inc.) proposed systems and processes for operating an intelligent automated assistant to provide a set of predicted responses. Cluster 4 and 5 focus on linguistics. Patent US20200327284A1 (assignee: ServiceNow Inc.) in cluster 4 proposed an agent automation system, which has processor that is configured to assign respective word vector to nodes and encodes semantic meaning of word or phrase represented by nodes. The system generates an annotated utterance tree by using a combination of rule-based and ML-based components, wherein an annotated utterance tree represents a syntactic structure of the utterance, and nodes of the annotated utterance tree include word vectors that represent semantic meanings. The annotated utterance tree is used as a basis for intent or entity extraction. Patent EP3111338A1 in cluster 5 also used automated text annotation for the construction of NLU

Table 3: Top 10 words and 2-gram phrases in each cluster.

Cluster	Size	Top 10 words and 2-gram phrases
1	20	Assistant, automate, user, input, language, natural, client, human, computer, processor Automate assistant, natural language, automated assistant, virtual assistant, assistant client, automate summarization, human computer, computer dialogue, input corpus, telephone call
2	27	Engine, language, natural, user, medical, code, processing, billing, generate, clinical Natural language, medical billing, billing code, language processing, language understand, patient encounter, clinical patient, free text, question answer, processing engine
3	45	User, request, response, language, natural, processing, action, query, input, generate Natural language, user request, action structure, language processing, language request, response user, speech input, dynamic training, computer readable, request text
4	31	Word, language, natural, phrase, computer, processing, plurality, sentence, processor, clause Natural language, target word, language processing, word clause, word phrase, neural network, input question, numeric code, program instruction, user interface
5	46	Text, language, natural, processing, user, processor, process, semantic, information, computer Natural language, language processing, language text, language understand, text interest, input text, semantic segment, information processing, touch operation, text natural
6	52	Plurality, language, entity, natural, generate, computer, associate, processing, document, name Natural language, name entity, language processing, computer readable, language input, cluster classification, reduced aggregation, flow diagram, machine learn, neural network
7	64	User, interface, language, NLP, natural, display, input, query, information, computer Natural language, user interface, language processing, user input, user query, graphical user, processing NLP, voice apparatus, real time, computer readable
8	30	speech, user, input, recognition, processing, language, determine, computer, audio, natural Natural language, speech recognition, speech processing, language processing, input audio, speech input, computer implement, user profile, language understand, automatic speech
9	52	Input, user, intent, language, natural, determine, processor, generate, NLU, computer Natural language, language input, user input, user intent, voice input, language understand, computer readable, input determine, transitory computer, user interface
10	35	Communication, user, language, natural, input, interface, voice, computer, processor, call Natural language, voice input, language processing, communication interface, input communication, text communication, voice communication, user input, phone call, communication channel
11	27	Application, user, language, natural, NLU, input, computer, associate, plurality, processor Natural language, online application, speech word, user online, language processing, language input, part speech, language understand, dimensional vector, structured natural
12	35	Information, language, module, natural, user, entity, service, input, generate, obtain Natural language, touch screen, language understanding, language understand, language processing, object hovering, understanding module, component process, target conversation, question answer
13	44	Language, natural, program, processor, structure, computer, instruction, user, analysis, expression Natural language, program instruction, language expression, frame structure, language processing, computer readable, language understand, semantic structure, language story, computer program

grammars. Patent US10789426B2 in cluster 5 described a device for processing natural language text with the context-specific linguistic model. Patent US10304444B2 (assignee: Amazon Tech Inc.) applies NLU to the music field, which uses a hierarchical organization of intents and entity types, and trained models associated with those hierarchies, so that commands and entity types may be determined for incoming text queries without necessarily determining a domain for the incoming text. Although cluster 6 is mainly concentrated in the "model" domain, there are also many terms related to "named entity." A representative patent US10755046B1 (assignee: Narrative Science) describes an NLP system for conversational inferencing with four-step parsing process.

Cluster 8 focuses on speech recognition. Patents US10446147B1 and US20200118564A1 (assignee: Amazon Tech Inc.) describe a speech recognition system to provide a contextual voice user interface. Patents US9245525B2, US9741347B2, and US10049676B2 describe an interactive response system mixes HSR subsystems with ASR subsystems to facilitate overall capability of user interfaces. Patents US9245525B2, US9741347B2, and US10049676B2 describe an interactive response system mixes HSR subsystems with ASR subsystems to facilitate overall capability of user interfaces.

Cluster 9 mentions about NLU, in which patent US9761225B2 (assignee: Nuance Communications Inc.) is representative. In US9761225B2, a method for identifying and resolving anaphora in multimodal conversational dialogue application for smartphone is proposed, in which multiple NLU interpretation selection models may be generated. The NLU interpretation selection models may include a generic model and one or more specialized NLU

Table 4: Domains, subdomains, and terms.

Domain	Subdomain	Cluster	Key words/phrases
NLP	Cognition	3, 13	Action, expression, action structure, frame structure, semantic structure, automate summarization
	Named entity recognition	6	Entity, name, name entity
	Linguistics (syntactic semantic, morphology)	4, 5	Part speech, word, phrase, sentence, semantic, plurality, document, clause, expression, target word, word clause, semantic segment, semantic structure, language expression
	NLU	9, 11, 12	NLU, user intent, user request, language understand, language understanding, understanding module
	Response	3, 12	Generate, question answer, response user
	Speech recognition	8	Speech, voice, audio, speech recognition, voice input, speech input, speech processing, input audio, speech word, automatic speech
Model	Model	6	Engine, machine learn, program, neural network, cluster classification, processing engine, dynamic training, dimensional vector, reduced aggregation
System	User interface	7, 11	Human, display, call, user interface, user input, touch operation, online application, phone call, telephone call, user online, graphical user, real time, human computer, touch screen, object hovering
	Medium	9	Computer, voice apparatus, computer readable, voice input, transitory computer,
	channel, communication	10	Communication, communication channel, communication interface, text communication, voice communication
Applied scenarios	Personal	1	Service, assistant, automate, client, automated assistant, virtual assistant, assistant client
	Medical	2	Medical, billing, clinical, medical billing, billing code, patient encounter, clinical patient
Skip	Skip	All	User, NLP, natural, language, interface, computer, analysis, query, recognition, process, processor, processing, structure, program, code, application, request, response, input, obtain, associate, service, instruction natural language, language processing, processing NLP, language input, input text, text interest, flow diagram, language text, determine, computer program, language input, information, input determine, input corpus, computer dialogue, free text, user query, language input, language request, input question, text natural, module, user input, user profile, component process, language story, structured natural, program instruction, frame structure, information processing, numeric code, computer implement

interpretation selection models, and each of which may be specific to a particular set of NLU interpretation type. Semantic reranking mechanism is applied in this method. Cluster 11 also mentions about NLU capability and focuses more on the follow-up actions, which are more related to "system" domain. Cluster 12 focuses on knowledge extraction in NLU. The representative patent is US10762113B2, which uses conversational knowledge graphs in virtual assistants to process natural language input, which involves receiving natural language queries from users at the virtual assistant's NLU system. Cluster 13 also belongs to cognition subdomain. Patents US9965461B2, US9594745B2, US9569425B2, and US20140249801A1 in cluster 13 (assignee: The Software Shop Inc.) describe the method for improving efficiency of syntactic and semantic analysis.

"Model" domain, concentrated in cluster 6, has no subdomain, and the number of key words is relatively low. The possible reason is that since neural networks are mainly mathematical algorithms and computers are only the carriers of mathematical operations, they cannot contribute to the technology themselves. In this case, what field the close integration of technologies and functions are in has come an important basis for judging technicality. If AI is only used to analyze business data, and technical problems are not solved, it is likely to be regarded as having no technical ideas, and it is difficult to overcome the nonpatent reasons by applying for repetition or amendment [42]. Algorithm-related patents must be combined with hardware-related terms as the carrier of the algorithm. This also explains the reason why cluster 9 contains a large number of "nontransitory computer readable device" vocabulary. The representative patents in cluster 6 are US10748526B2, US10747958B2, and US10733375B2.

"System" domain contains user interface, medium, and communication or channel subdomains, in which four clusters, cluster 7, 9, 10, and 11, are distributed.

As for "applied scenarios," concentrated in cluster 1 and 2, terms such as "virtual assistant," "medical," and "billing" are found. In cluster 1, three patents assigned to Google LLC are representative for virtual assistants in "personal" subdomain. Patent US20200320136A1 proposes a method for using distributed state machines for human-to-computer dialogues with automated assistants to protect private data.

Patent US20200050788A1 describes a system for assembling responses from remote automated assistants. Patent KR2020131299A proposes a method for generating Internet of things-based notification by automated assistant client of client device. In cluster 2, three patents assigned to Nuance Communication Inc. are representative for medical billing and coding in "medical" subdomain. Medical billing and coding are two closely related aspects of the modern health care industry. Both practices are involved in the immensely important reimbursement cycle, which ensures that health care providers are paid for the services they perform [43]. Patent US20170323060A1 describes a system with a graphical user interface (GUI) and an NLU engine to automatically derive one or more engine-suggested medical billing codes. Patent US10319004B2 proposes techniques to deal with the overlapping codes derived by the NLU engine, and patent US10754925B2 proposes a method for training NLU engine, involves providing training data in form of free-form text, corrections, and finalized sequence of medical billing codes.

Three domains were found from the clustering results. It is particularly important to emphasize that the composition of natural language-enabled chatbot mostly relies on the three domains, NLP, model, and system. Since most of the related patents contain these three parts at the same time, it is difficult to determine the exact belonging domain for each patent and also meaningless.

3.3. Topic Modeling. According to the ontology construction process (see Figure 1), search query, corresponding result, and topics founded in each domain are illustrated in Table 5. For domain NLP and system, DI smart search is applied, while CTB (claim/title/abstract) strategy is applied for domain model and applied scenarios. Table 6 illustrates the keywords of each topic.

3.4. Applied Scenario Topic Modeling. This research hopes to find the application field of NLP chatbot, but a lot of experts are describing natural speech-related technologies or the system framework of conversation management, which are not discussed in this section. This research mainly divides the application scenarios into engineering applications and e-commerce applications. It can be found from the patent search results that natural language-enabled chatbot is widely used in the field of e-commerce, while the application on the engineering side is difficult to find. 44 patents are reviewed manually and classified to certain topic or scenario. These patents with respect to the applied scenario are listed in Table 7.

Here are some patents in topic of e-commerce. Patent US20170323060A1 describes a system for facilitating automated natural language understanding for medical documentation of patient, which has processor for presenting set of medical billing codes for user review in graphical user interface (GUI) before finalizing coding of encounter. Patent KR2020000621A describes a conversation system for grasping user attention during various situations in a vehicle by using a mobile device. The system has a storage unit for storing situation information collected from a vehicle. A dialogue management module obtains a factor value of action factor used to perform an action corresponding to a dangerous situation when an input processor obtains an action corresponding to the starting situation from the storage unit. An input processor generates a dialogue to perform the action corresponding to the dangerous situation by using the factor value of the acquired action factor while obtaining the action corresponding to the dangerous situation and generates a conversation message. A result processor generates a conversation response corresponding to a delivered starter message. Patent US10223934B2 proposes a method for monitoring and analyzing language environment, vocalization, and development of key child, which provides metrics associated with key child's language environment and development in a relatively quick and cost-effective manner. The proposed method is used to promote improvement of the language environment and key child's language development and to track development of the child's language skills. Key child's language environment and language development are monitored without placing artificial limitations on the key child's activities or requiring third party observer.

Here are some patents in topic of engineering. Patent JP06792132B2 defines an information-processing apparatus, which is used in the manipulator control system and NLP system and can be performed with high versatility. The information-processing apparatus has processing module groups, and each of which is equipped with several processing modules with specific processing capabilities. These processing modules have a neural network with a hierarchical structure. The information is processed by sending and receiving the information signal of the processing module in several interhierarchical structures. Patent CN111267097A proposes a natural language-based assisted programming method for industrial robots, involves parsing language instructions, matching parsing result, and combining coordinates output to generate final robot auxiliary code. The multiattention mechanism model adopted by the method improves the recognition accuracy and solves the problem that the current method cannot accurately recognize objects in an industrial environment. Modular programming technology solution simplifies engineers programming complexity and effectively improves development efficiency. Patent US10843080B2 describes a system for facilitating automated program synthesis from natural language. The system allows a user to be more comfortable and familiar with grammatical requirements for forming a proper sentence in native language as opposed to memorizing rules or required constructs for a potentially complicated programming language. The system employs fuzzy grammar matching to reduce complexity, while slightly trading off complexity for accuracy. The system allows the user or developer to examine to express an idea in a different manner to better reflect user an original intent. Patent DE102018212503A1 defines communication and control systems, which has control devices for operating machine based on software communication chatbot, for filling beverage in bottling plants. The chatbot recognizes a voice input

Table 5: Topics in each domain.

Domain/method	Query		
	Search type	Input patent size	Topics
NLP/LDA	SSTO = ("natural language processing" "linguistics" "natural language generation" "natural language understanding" "speech recognition")		
	Smart	570	Linguistics, conversation, speech recognition, knowledge
Model/LDA	CTB = ((chatbot) or (automated adj conversation* adj interface*) or (chat* ADJ system*) or (natural adj language*) or (nlp*)) AND CTB = (((deep ADJ learning) or (machine ADJ learning) or (neural ADJ network))) AND DP> = (20200101) AND DP< = (20201130)		
	CTB	2,535	Features, voice device, question answer, classification, graph, automatic service
System/LDA	SSTO=("natural language processing" "natural language understanding" "NLP" "NLU" "chatbot" "automated conversational interface") AND SSTO = ("user interface" "medium" "communication" "channel" "immersive technology" "computer vision")		
	Smart	534	User interface, dialogue management, infrastructure
Applied scenarios/ manual	CTB = (((chatbot*) or (conversation*) or (diolog* adj system*))) AND CTB = (((natural adj language*) or (nlp*))) NOT DC = ((T))		
	CTB	31	Engineering, e-commerce

Table 6: Keywords for each topic.

Domain	Topic	Keywords
NLP	Linguistics	Personality, AI, discourse, syntactic
	Conversation	NLU, semantic, NLG, intent
	Speech recognition	Audio signal, processor, channel
	Knowledge	Entity, ontology, semantic, cognitive identification
MODEL	Features	semantic, vector representation, image recognition
	Voice device	Storage, server, control module
	Question answer	Pair, retrieval, RNN
	Classification	Segmentation, convolutional, encoder
	Graph	Entity, ontology, intent
	Automatic service	Call, recommendation
System	User interface	Portable, network, wireless, digital
	Dialogue management	NLG, processor, ML
	Infrastructure	Channel, cloud, communication

Table 7: The manual induction result by applied scenario.

Topics	Scenario	Publication number		
E-commerce	Medical	US20170323060A1	WO2020061562A1	US10679345B2
		US20200185102A1	JP2020518047A	US20200027535A1
	Health	CN111612752A	US10754925B2	US10319004B2
		CN109591024A	US10748644B2	
		KR2020000621A	WO2020069517A3	US10752212B2
	Driver assistant	CN111145731A	US20200216089A1	US10543931B2
		US10573299B2	US20200135183A1	EP3606797A2
	Exercise	KR2173553B1	US20200114207A1	
	Education	US10223934B2		
	Emotion	US10579742B1	CN111312394A	
	Smart home	IN202041050057A	CN110654738A	
	Customer service	CN108282587B	IN201821029643A	CN111902878A
	Smart assistant	US10748526B2	US10747958B2	US10733375B2
		EP3753017A1		
	Entertainment	EP3566399A4		
Engineering	Robot	CN111645073A	CN111267097A	JP2020526402A
		JP06792132B2	US20200306958A1	
	Programming	US10843080B2		
	Manufacturing	CN107632845B	DE102018212503A1	
	Quality control	WO2020181365A1		

and a text input by an operator to output or display information about an operating state of the machine. The systems realize production conversion of energy in an automatic manner and order completion in a rapid manner and improve media efficiency and scheduling efficiency. Patent WO2020181365A1 proposes an apparatus for 360-degree assistance for quality control system scanner with mixed reality (MR) and ML technology. The apparatus has an optical sensor, a display, and a processor to receive diagnostic information from a server related to a field device in an industrial process control and automation system. The processor identifies an issue of the field device based on the diagnostic information, detects, using the optical sensor, the field device corresponding to the identified issue, and guides, using the display, a user to a location and a scanner portion of the field device that is related to the issue. The processor provides, using the display, necessary steps or actions to resolve the issue, and connects, using a cloud server, a user to get modules of installation, commissioning, AMC, and training for a QCS as per the selected person.

3.5. Ontology. In this section, the ontology map of NLP chatbot is drawn based on the previous outputs. A four-level ontology includes subject, domains, topics, and key phrases in a top-to-bottom sequence. Under the subject of NLP chatbot, the domains are NLP, model, system, and applied scenarios. The third level has the topics under each domain. For NLP domain, there are speech recognition, linguistics, conversation, and knowledge. For domain of model, topics are feature, graph, voice device, question answering, classification, and automatic service. For domain of system, the topics are infrastructure, dialogue management, and user interface. For applied scenarios, e-commerce and engineering are the two main topics. The fourth level has the key phrases under each topic. It is noticed that some key terms are shared by multiple topics. The ontology map of NLP chatbot is shown in Figure 2.

4. Patent Macro Trend Analysis

Related patents are searched by entering keywords related to NLP and chatbots on the DI database, and patent management map analysis is conducted (see Table 8). From 2011 to 2020, totally 21,834 individual records or 12,840 DWPI families are published. Patent family refers to the collection of patents applied for in different patent offices for the same invention. DWPI has a stricter definition. Each patent in the same DWPI patent family must have exactly the same priority as other patents in the family. The analysis of this section is mainly based on DWPI families. The following term "patents" refers to "DWPI families" unless otherwise specified.

Since 2017, 10,480 patents have been published, accounting for 82% of the total 12,840 patents in the past decade. Furthermore, since 2019, 8,099 patents account for 62%. From the perspective of the annual growth rate of the number of patents, the number was a high 44% in 2014, but returned to 6% in 2015, which is the lowest number in the past decade. However, starting in 2016, the annual growth rate has increased sharply until it reaches a peak of 105% in 2019, and it then falls back to 66% in 2020. Whether the decrease in the number of 2020 is related to the impact of COVID-19 is unknowable, but this may be a signal that implies that the technology related to natural language-enabled chatbot may have gradually matured.

However, a single reduction in quantity cannot lead to any conclusions unless supported by more other data or evidence. IPC is a standard taxonomy developed and administered by WIPO for classifying patents and patent applications, which covers all areas of technology and is currently used by the industrial property offices around the world. From the annual number of patents with IPC analysis, to 2018, all The IPC classifications have been covered. In other words, among the 8,099 patents in 2019 and 2020 that accounted for 62% of the number in the past decade, no new technology has been produced.

Top 6 4-character IPCs, with a number of patents that greater than 1,000, are G06F (electric digital data processing), G06N (computer systems based on specific computational models), G06Q (data processing systems or methods), G10L (speech analysis or synthesis; speech recognition; speech or voice processing; speech or audio coding or decoding), H04L (transmission of digital information), and G06K (recognition of data), each in which has a number of 8,870, 3,144, 2,413, 2,176, 1,364, and 1,258 patents, respectively (see Figure 3). It should be noted that the total proportion can exceed 100%; that is, the summation of these number can be greater than 12,840, because a patent can be classified as multiple IPC codes.

G06F's patents accounted for 8,870 of 12,480 patents. Therefore, the complete IPC classification in G06F was further explored. Among the top 10 IPCs listed (see Figure 4), 2,295 patents are classified in G06F 17/27 (for automatic analysis, parsing, orthographic correction, etc.). The second largest class is G06F 17/30 (for information retrieval and database structure). It is worth noting that the 3rd and 4th classifications (G06N 3/04 and G06N 3/08) represent the interconnection topology architecture and learning method, respectively. G06F focuses on data processing procedures, while G06N emphasizes system structure. G06F and G06N domain classifications represent the key technologies for implementing the main modules of complex natural language-enabled chatbot systems. In addition, G10L 15/22, ranked 9th, is about programs used in speech recognition for human-machine dialogue.

In addition to statistics on the number of patents, the fluctuations in the number in recent years are also worthy of attention. Based on the annual growth rate of all patents, when the growth rate of an IPC is higher than average, it represents greater momentum; conversely, when the growth rate of an IPC is lower than average, it may imply that the technology has entered the mature stage early. The four 4-character IPCs with the largest number were selected for this analysis (see Figure 5).

G06F has an overwhelming 69% of total patents, but its annual growth rate is much inferior to the average annual growth rate. In 2014, the total number of patents related to

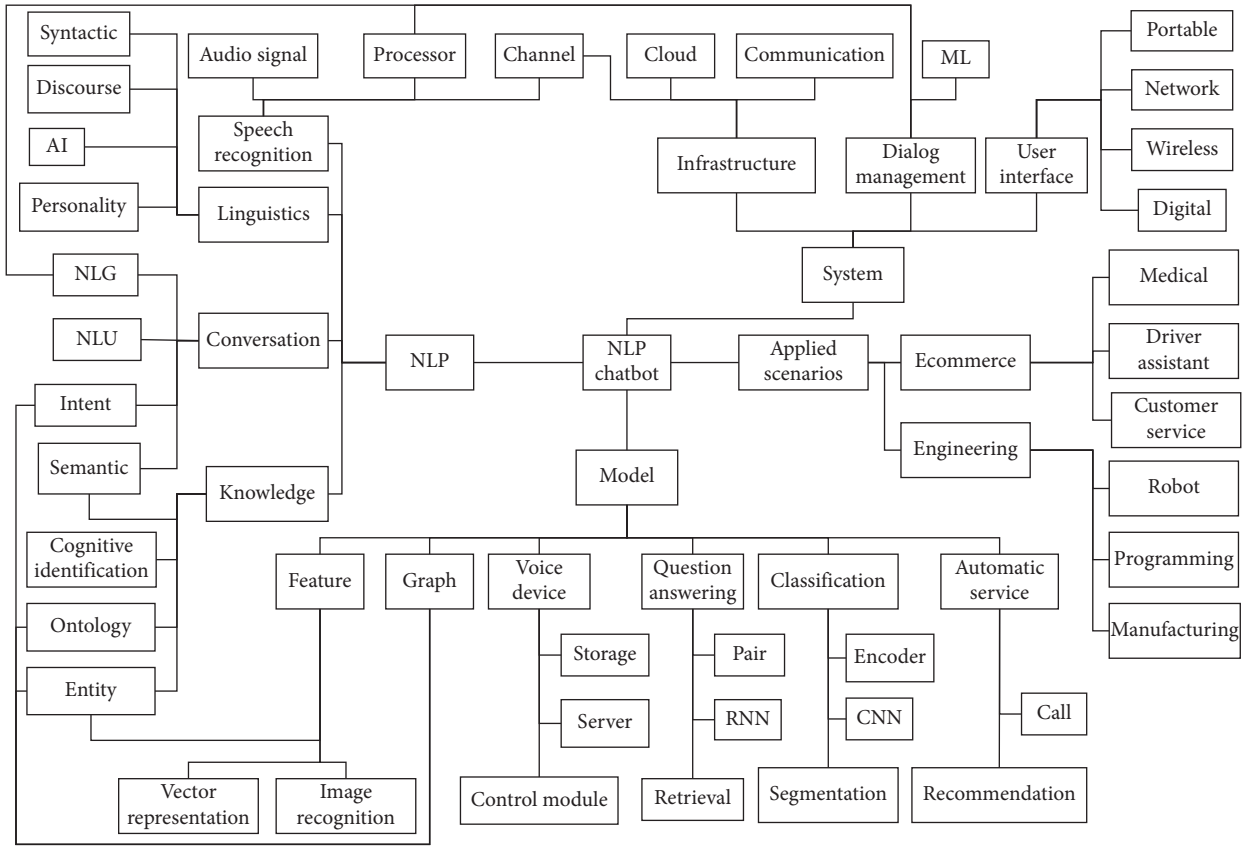


Figure 2: NLP chatbot ontology.

Table 8: Search query for patent management analysis.

Search type	DI query	Result
Claim/title/abstract	CTB=((chatbot) or (automated adj conversation* adj interface*) or (chat* ADJ system*) or (natural adj language*) or (nlp*)) AND CTB=(((ontology) or (named ADJ entity ADJ recognition) or (deep ADJ learning) or (machine ADJ learning) or (neural ADJ network) or (speech ADJ recognition) or (feature*) or (lstm) or (gated adj recurrent adj unit) or (transformer) or (BERT) or (GPT*) or (rectifier) or (RELU) or ("speech%" ADJ "generat%") or (cloud ADJ computing) or (voice ADJ activity ADJ detection) or (voice ADJ over ADJ Internet ADJ protocal) or (bandwidth) or (human ADJ computer ADJ interaction) or (VUI) or (GUI) or (user ADJ interface) or (immersive ADJ technolog*) or (virtual ADJ reality) or (augmented ADJ reality) or (mixed ADJ reality) or (force ADJ touch) or (3D ADJ touch) or (robotic ADJ process ADJ automation) or (communication ADJ system))) AND DP>=(20110101) AND DP<=(20201231);	12,840 DWPI families

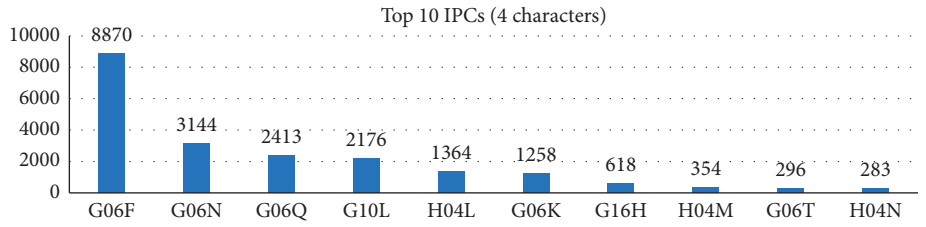


Figure 3: Top 10 IPCs (4 characters).

natural language-enabled chatbot rose sharply by 44.37%. The growth rate of G06F in that year was only 41.40%, which was slightly lower than the average. Since 2016, during the period of rapid growth in the number of patents, the growth rate of G06F has not been outstanding. Even when the average growth rate reached a peak of 104.49% in 2019, G06F

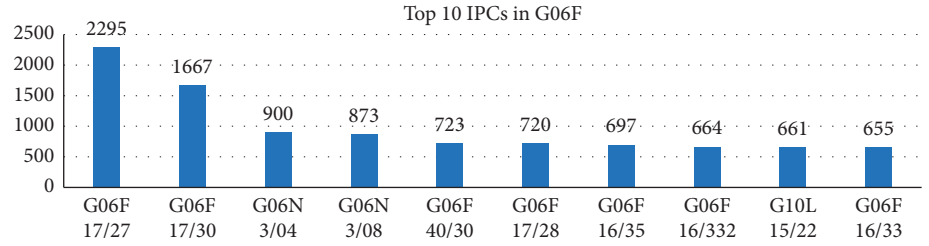


Figure 4: Top 10 IPCs in G06F.

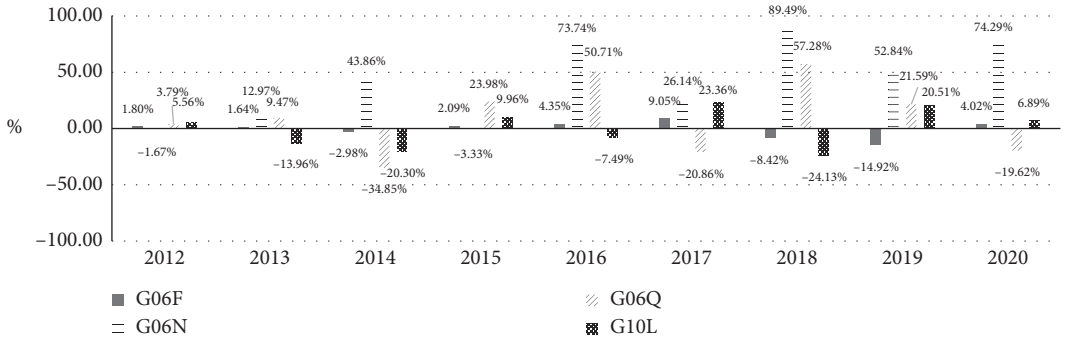


Figure 5: Annual patent growth rate under top IPCs.

was 14.92% less than the average. By contrast, the annual growth rate of G06N is amazing. In 2014, it was 43.86% higher than the average, and from 2016 to 2020, the annual growth rate was 73.74%, 26.14%, 89.49%, 52.84%, and 74.29% higher than the average, respectively. G06Q and G10L fluctuate up and down in average annual growth rates and have not yet shown a clear trend.

In general, the average annual growth rate began to slow down after reaching a peak in 2019 after rapid growth, no new IPC appeared after 2018, and all of which indicate that the development of natural language-enabled chatbot has entered a mature stage. It is worth noting that the patents related to only reading G06N are still growing rapidly.

Assignee analysis helps to find the main players in the market, which are all technology giants from the results. The number one IBM has 1,358 patents, which is more than the total number from the second to the tenth. The well-known technology giants Apple Inc and Facebook Inc are ranked 16th and 17th, respectively. Although they are not in the top 10, they are also listed in the table due to their influence (see Table 9).

IBM's patents began to grow rapidly in 2016, when IBM's patents were concentrated in the two categories of G06F 17/30 and G06F 17/27, showing that IBM focused on information retrieval and grammar analysis in NLP. In 2019, the number of patents of Microsoft, Amazon, Accenture, and Univ Kunming Science and Tech began to grow significantly. In addition to G06F 12/27, Amazon and Microsoft use speech recognition technology based on natural language models in human-machine dialogue, which is mainly reflected in the two IPCs G10L 15/18 and G10L 15/22. In 2020, the number of patents of Google, Samsung, and Baidu increase rapidly at the same time. In addition to the two categories of G10L 15/18 and G10L 15/22 related to speech recognition in 2019, both Google and Samsung have more patents appearing in G06F 3/16, which focuses on the conversion between speech and digital information. On the other hand, Google and Baidu applied for many patents on G06N 3/08, which are the computer system based on learning methods. In addition, Baidu also has a large number of patents on G06F 40/30 for semantic analysis. Google and Baidu are both Internet service companies that started as search engines, and Google and Samsung are also close partners in the android camp. The highly increasing number of patents assigned to these three companies, which are quite close to the end user, might imply the maturity stage and mass application in this technology field. From the IPC distribution of Apple Inc.'s patents in 2019 and 2020, it can be seen that its patents are highly concentrated on speech recognition-related G10L 15/18, G10L 15/22, and G06F 3/16, which are similar to Google. Google and Apple coincidentally began to cut into a large number of patents in the field of speech recognition, speech, and digital information conversion in 2019. The clues can also be seen from their products. The Google Nest Mini launched in November 2019 and the Apple HomePod launched in August 2019 show the development path from smart speaker to smart home. With the maturity of natural language technology and IoT, the use of natural language to control objects around life will gradually replace the previous method of operating through buttons or operating with limited system interfaces. When other companies focus on deepening NLP-related technologies or developing speech recognition applications, Facebook Inc. has paid more attention to electric communication technique, including H04L 12/58 and H04 29/08.

Table 9: Top 10 assignees.

Top	Assignee	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
1	IBM	12	14	22	24	26	89	129	231	313	498	1,358
2	Microsoft Technology Licensing LLC	0	0	0	0	12	14	23	39	72	125	285
3	Amazon Tech Inc.	0	2	0	0	8	16	15	35	64	57	197
4	Google LLC	0	1	0	0	0	0	0	15	36	133	185
5	Samsung Electronics Co. Ltd.	2	4	4	3	5	6	8	19	27	74	152
6	Nuance Communications Inc.	4	4	9	14	12	12	17	15	15	23	125
7	Accenture Global Solutions Ltd.	0	0	0	0	0	0	2	7	37	77	123
8	Beijing Baidu Netcom SCI & TEC	0	0	0	0	0	0	1	6	10	92	109
9	Microsoft Corp	21	21	14	17	8	2	4	3	2	2	94
10	Univ Kunming Science and Tech	0	1	0	0	1	3	1	15	47	23	91
16	Apple Inc.	0	1	0	3	1	1	6	10	19	36	77
17	Facebook Inc.	0	0	0	2	1	3	5	4	19	38	72

The two IPC codes represent message switching systems and transmission control procedure in network communication, respectively.

5. Technology Function Matrix

A Technology Function Matrix (TFM), which investigates the corresponding relation between technologies and functions on patent amount, is a critical approach for patent data analytics. The domain of NLP, model, and system, which is introduced before in Section 3.2.3, are used to form the TFM. The construction process of TFM is described as the following. A well-constructed ontology is defined before, from which technology and function terms can be defined, and patents can be collected by the search query set according to the ontology. Next, each patent is visited iteratively to count if it matches each technology and function. By doing this, a TFM can be constructed.

This research uses the TF-IDF-based TFM automatic construction method. After defining the technologies and functions, an unstructured text description that best represents each technology or function must be prepared. These text descriptions are transformed into a set of vectors through unsupervised learning, which acts as an agent for each technology or function. Then, specific fields are selected from each patent, converted into a vector, and compared with each technology and function through similarity, and a threshold is used to determine whether the patent can be classified as the technology or function. Thus, the text description of each technology or function is very important. Sections 5.1 and 5.2, respectively, explain the technologies and functions selected in this study, followed by the TFM result in Section 5.3. After that, the domain of applied scenarios is added to form the three-dimensional matrix, which is called A-TFM and is introduced in Section 5.4.

5.1. Definition of Technology. 13 TFM technologies, listed below in Table 10, are defined according to domain of NLP, model, and system. The description of the similarity compared with the patent text is extracted from Wikipedia. Speech recognition, NER, NLU, and NLG are technologies in the domain of NLP. Feature engineering, RNN, CNN, and transformer are of model. And speech-generating device, cloud computing, voice activity detection, human-computer interaction (HCI), and immersive technologies are of system.

5.2. Definition of Function. Nine TFM functions, which are information extraction, dialogue management, context prediction, recommendation system, algorithm efficiency, automated control, communication, user experience, and virtual assist, are listed in Table 11. The description of the similarity compared with the patent text is extracted from Wikipedia and other web resources.

5.3. TFM Result. For finding emerging trend of natural language-enabled chatbot, year 2020 patents are used as the source for TFM. The 13×9 TFM result is obtained through the automated process described before (see Table 12). Transformer is a DL language model, developed in 2017, widely used to process natural language tasks. The patents related to transformer technology and prediction function are the highest number, which means transformer is a mature technology and be widely applied for context prediction. In terms of technologies (row), transformer and speech-generating device are the main technologies of the current market and have a positive impact on almost all functions. In terms of functions (column), automated control function is more widely used than others. For instance, speech recognition and speech-generating device are for increasing the pipeline of the control system. In addition, the NLP domain technologies mostly relate to information extraction, dialogue management, and prediction, such as the improvement of NLU and NLG can enhance the system's ability to identify users' intent. Last, the system domain technologies mostly concentrate on communication, user experience, and virtual assistant. For

Table 10: TFM technologies.

ID	Domain	Technology
T01	NLP	Speech recognition
T02	NLP	Named entity recognition
T03	NLP	Natural language understanding
T04	NLP	Natural language generation
T05	Model	Feature engineering
T06	Model	Recurrent neural network
T07	Model	Convolutional neural network
T08	Model	Transformer model
T09	System	Speech-generating device
T10	System	Cloud computing
T11	System	Voice activity detection
T12	System	Human-computer interaction
T13	System	Immersive technologies

Table 11: TFM functions.

ID	Function
F1	Information extraction
F2	Dialogue management
F3	Context
F4	Recommendation system
F5	Algorithm efficiency
F6	Automated control
F7	Communication
F8	User experience
F9	Virtual assistant

instance, adopting immersive technologies can enhance user experience or the development of cloud computing makes portable devices handling complex tasks. Therefore, a lot of virtual assistants are developed to assist people for a convenient life, such as intelligent drive assistant. Next, the interaction of technology and function and its related patents are explored to find emerging technologies or applications.

5.3.1. Speech Recognition (T01). The most applied function of speech recognition is information extraction (F3). Accuracy of speech recognition is the key to determining whether it can be applied to the commercial field, and good information extraction ability is a necessary condition. Although speech recognition technology has gradually matured, there are still a large number of patents in this field for better recognition capabilities and information extraction capabilities.

Google LLC's patent US10431206B2 uses the hierarchical recurrent neural network (HRNN) structure handles the task of multiaccent speech recognition. Patent CN110033766A proposes a complex multiple deep neural network architecture, including single layer of one-way RNN model, binary bidirectional RNN model, and binary bidirectional LSTM (BiLSTM) model and other network structure, in pursuit of faster speed and less energy consumption. Patent EP3497630B1 uses CNN architecture, which allows better signal propagation and long-range dependency learning, thus improving output quality.

In addition, speech recognition and automated control functions (F6) are combined with each other to form the application of speech-driven automated control. When receiving speech data from the client, speech recognition and NLU model stored in the cloud are used to interact with other devices in the cloud space, such as unmanned aerial vehicles (UAVs), robots, augmented reality (AR), and virtual reality (VR) devices, through AI modules and 5G network technology.

5.3.2. NER (T02). In order to improve the accuracy of NER, preprocessing is very important. Patent CN110990525A proposes a sentiment-based information extraction method that achieves good performance in the field of financial sentiment information extraction through preprocessing and feature extraction modules. Data labeling and feature engineering are the two main steps in preprocessing. Patent CN111783466A proposes a named entity recognition method for Chinese medical record field, in which the label uses two-layer conditional random field (CRF) classification to determine the final output label thus improving the accuracy of NER and reducing the time consumed by training. There is similar research in literature studies. In view of the insufficient representation of potential features of Chinese characters, Han et al. [44] uses the BiLSTM network to learn the internal strokes and radical semantic information of Chinese characters and combines with the BiLSTM-CRF model to construct an adaptive multifeature fusion embedded CNER model. In addition, patent WO2020167558A1 proposes a dynamically trained model of named entity recognition over unstructured data, which defines entity labels for specific domain knowledge ontology, and uses these entity labels to identify the relationship between unstructured documents and domain knowledge. Patent CN111737969B proposes a resume analysis method based on a DL model, which combines NLP, OCR, and named entity recognition technology. This method first performs feature modeling on the resume. After the model training is completed, the key information is classified and the category mapping model is set, so that the parser can read it like a human and improve the overall analysis effect.

5.3.3. Transformer Model (T08). The transformer model is widely used to improve the accuracy of the information extraction function (F1). Patent CN110941698A proposes a method based on the bidirectional encoder representation on BERT CNN, which generates rich contextual semantic information of word vectors, thereby effectively supporting service similarity calculation to find the most accurate target service, and achieving accurate retrieval of target services.

As for dialogue management function (F2), patent CN111274362A proposes a dialogue generation method based on the transformer architecture, which involves obtaining a vectorized representation of words, and generating a reply based on a comprehensive semantic vector and a copy mechanism, which is used to solve the NLG based on background domain knowledge dialogue. Patent US20200372341A1 proposes a pipelined natural language

Table 12: The TFM result.

			F1	F2	F3	F4	F5	F6	F7	F8	F9
T01	NLP	Speech recognition	703	673	883	343	307	661	452	484	630
T02	NLP	Named entity recognition	948	412	484	301	339	852	134	318	217
T03	NLP	Natural language understanding	809	503	514	195	177	627	87	249	188
T04	NLP	Natural language generation	989	724	827	308	348	759	105	378	272
T05	Model	Feature engineering	571	377	306	436	372	470	136	299	190
T06	Model	Recurrent neural network	569	386	613	257	244	332	190	110	161
T07	Model	Convolutional neural network	317	251	446	287	327	208	149	98	144
T08	Model	Transformer model	1,048	995	1,348	514	422	714	308	452	448
T09	System	Speech-generating device	1,189	1,141	1,123	622	384	1,006	792	998	963
T10	System	Cloud computing	341	422	237	358	213	465	487	452	458
T11	System	Voice activity detection	309	377	241	283	120	379	780	557	705
T12	System	Human-computer interaction	685	858	509	512	260	626	803	909	850
T13	System	Immersive technologies	307	417	211	272	123	211	439	533	420

question answering system based on the BERT model, which involves receiving an input text of a natural language question and provides an answer to the natural language question considering context.

The transformer model is used in context (F3) function to improve the accuracy of NLP. Patent CN110737764A proposes a method for generating personalized dialogue content based on a multiround dialogue model. The transformer model effectively learns the dialogue sequence relationship between natural languages, can predict the generated content to reduce the probability of replying commonality, and increase the diversity of dialogue content. Patent CN111708882A proposes a method for complementing missing Chinese text information based on transformer encoder. This method starts from manually preprocessing Chinese text documents, dividing the text into a large number of short sentence corpora, and converting it into the smallest unit of BERT vector. Since the purpose is to find out the missing words and sentences in the article, the training method is to randomly generate noise to hide the words in the complete article to create the effect of the omission. Conversely, in order to be able to fill in the missing words, the model must have text generation capabilities. Through repeated information deletion and generation procedures, Chinese natural language processing task accuracy is further improved.

5.3.4. Speech-Generating Device (T09). Speech-generating device is highly related to the three functions of information extraction (F1), dialogue management (F2), and context (F3), with 1,190, 1,141, and 1,123 patents, respectively. The speech recognition technologies of T09 and T01 are also highly related, but the classification of T09 in the "system" domain means that the description of this technology is more focused on the hardware or system framework, so that for T09, F1, F2, and F3. The gap between is blurred. From these large numbers of patents, it can be found that with the maturity of Internet technology and mobile devices, the past information retrieval systems have begun to be replaced by chatbots. However, when NLP technology is not yet mature, rule-based chatbots cannot exert influence. However, as NLP technology and speech recognition technology mature, speech-generating devices have also developed rapidly and combined with chatbot applications. Task-oriented retrieving systems began to be replaced by speech query systems. Patent CN110111766A claims a multifield multitask system, which solves the problem of the multidomain multitask switching in the dialogue system. The complex multitask dialogue system integrates a speech recognition module, a domain confidence state tracking module, dialogue managing module, an NLG module, and a speech synthesis module to realize the capability that semantic level information can be shared between each domain. Patent JP2020098308A proposes a voice inquiry system for information provision, in which each of chatbot servers and smart speaker operation server use the DL model, accept a spoken question, infer, and output the corresponding answer in spoken speech.

The next step after reaching the speech query system is speech-driven remote control. 1,006 patents related to automated control function also support this idea. Patent US10748529B1 (assignee: Apple Inc.) proposes a voice-based digital assistant for use with home automation of voice activated controllable device, such as TV, speaker, or camera. The application of speech-driven automated control is not so uncommon, but they are focused on devices that do not have safety hazards, such as home-related devices. It also means that speech-driven automated control is still at the auxiliary stage and cannot replace existing functions. However, it is believed that one day people will hope that many functions that require physical contact can be replaced by voice control, and the first thing to overcome is noise. Since the sound is not specified, the device may receive unexpected sounds and trigger actions at any time. Therefore, a gateway may be required to avoid unexpected actions caused by noise. Patent US20140214414A1 proposes a

communication system for use in automatic speech recognition applications, which can transmit commands through wireless network to modify gateway's noise reduction processing state.

5.3.5. HCI (T12). When it comes to smart homes, in addition to speech control, there are more automatic control methods through HCI. Patent CN110932953A proposes s smart home control method and device, which can receive the user control command of the target home, login target start home residence in the target network, intelligently perform control, and return the result message back. This solution realizes the multihome for different manufacturers and different communication protocols for uniform control.

It is observed from TFM that HCI technology is widely used to improve user experience (F8), and there are 909 patents located in the interaction. Most people use chatbots to meet their needs, such as information retrieval or specific operational tasks. It is most important to be able to meet the needs of users in fewer conversations. Many patents also aim to reduce dialogue and improve dialogue efficiency, such as CN112015879A, CN110990594A, CN111488433A, and CN110827831A.

5.3.6. Immersive Technologies (T13) with Virtual Assistant (F9). In addition to the HCI methods of contact and voice, the use of gaze tracking to help virtual assistants more accurately grasp the text or dialogue paragraph the user is paying attention to is an emerging application.

5.4. A-TFM and TFM with Applied Scenarios. As mentioned in Section 3.4, the applied scenario factor is also a valuable part for analyzing patents. Therefore, this research utilizes the applied scenarios as the third dimension to construct a 3-dimensional matrix. As shown in Figure 6, the scale of node means the number of patents. *X*-axis means 13 technologies, *Y*-axis means 9 functions, and *Z*-axis means 7 applied scenarios. The source of this three-dimensional matrix is 50 patents which randomly collected from the source of the above TFM. "Personal" and "e-commerce" are the main applied scenarios of the current market. "Medical," "engineering," and "driver assistant" are applied scenarios still under development. Also, few patents related to "education" and "society" chatbots are found.

6. Discussion

Nine topics, including medical data, smart cities, IoT, data privacy, sustainable strategies, CRM, personalization, social media listening, and ML models, are identified as latent topics for future research based on data-driven strategies [14]. This research thoroughly investigates the application of chatbots by comprehensive patent-mining process and claims the consistency between the findings of this study and the above results. Thus, the effectiveness of proposed analysis is justified.

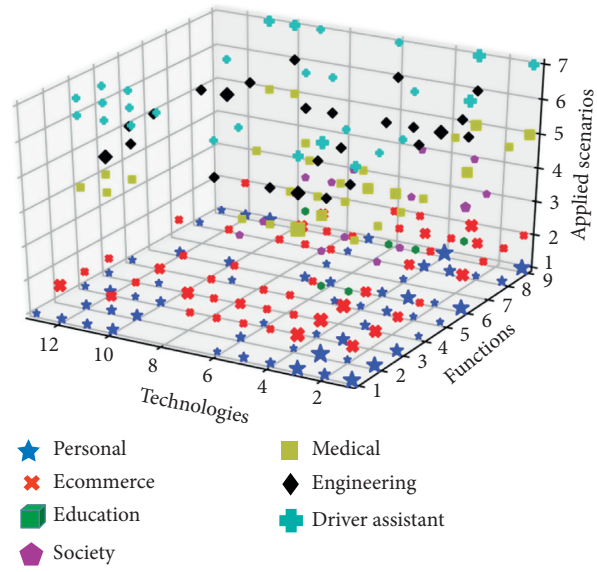


Figure 6: The three-dimensional matrix.

6.1. Knowledge Graph. AI makes huge progress; algorithms are rapidly improving, managing massive amount of data; however, it still is not knowledge-driven technology. The knowledge behind the natural language-enabled chatbot is very important for dialogue with humans. The early development of chatbot was mostly dominated by a single domain. It has been observed that more research has been directed towards open domain [45–49] and multidomain [50–52] in recent years. Single-domain chatbots are limited to accomplishing specific tasks, while multidomain or open domain chatbots can better meet the needs of smart assistants and even further provide people's companionship or social applications. With the development of 5G and cloud applications combined with social media, many social media, such as Telegram, Cortana, Slack, WeChat, Facebook Messenger, Google Assistant, and Siri, provide platforms that can easily build chatbots [53], making the transition of technology bottleneck shifting from simple single-domain chatbot system construction into complex integration of multidomain knowledge bases. The correlation between these two phenomena is hypothesized.

With the rapid development of the semantic web, a large amount of structured data has been provided in the form of a knowledge based on the web. Making these data accessible and useful to end users is one of the main goals of chatbots based on link data [54]. KG is considered to be a new AI technology trend, which originated from the basic principles of the Semantic Web and the construction of the knowledge base [55]. The novel KG-based framework is used in many chatbot applications. They combine the query language SPARQL of the resource description framework to quickly integrate the existing knowledge base.

Related patents in recent years have also focused on studying how the knowledge framework can improve the capabilities of NLU and integrating the KG into the knowledge base of chatbot. Patent US10733375B2 (assignee: Apple Inc.) provides a system and process for operating

intelligent automated assistants. This process is based on a knowledge framework and can improve the validity of NLU, analyze the mapping of domain attributes and words from the natural language input, then correspond to the data of the knowledge base according to the analysis results, and determine the output response results according to the ranking mechanism. Patent EP3362972A1 proposed a system for authoring visual representation for text-based natural language document. User interface is provided that contains a document area and thus enables to interactively generate the visual representation information that accurately depicts the underlying source text. The system generates a node graph of at least one of the parse trees, the entity information, or the relational phrase information and processes the document to determine relational phrase information indicating that the portion of the text includes a relationship to at least one of a subject, verb, or object in a sentence that includes the portion of the text. Also, the system generates another visual representation links the nodes and the relations. Patent WO2020160264A1 proposed a method of identifying relevant data sets using training models related to topics of interest, involving access to one or more sources, each of which contains information systems and related methods used to organize, represent, find, discover, and access data. The embodiment represents information and data in the form of a data structure called a "feature graph." The feature graph includes nodes and edges, where edges are used to "connect" nodes to one or more other nodes. The nodes in the feature graph can represent variables, that is, measured objects, features, or factors. The edge in the feature graph may represent a measure of the statistical association between a node and one or more other nodes that have been retrieved from one or more sources. The data set that represents or supports statistical correlation or measurement correlation variables is "linked to" form the "feature graph." Patent US10762113B2 (assignee: Cisco) proposes the use of conversational knowledge graphs in virtual assistants to process natural language input. After receiving the natural language query of the user, the method retrieves the contextual information of the conversational knowledge according to the intention and calls the back-end service accordingly and obtains the response after the service is performed. Finally, the response is translated into natural language and provided to the user. There are similar studies in literature studies. Zhong et al. [56] designed a cognitive information representation model based on the knowledge graph, which combines the perception information and semantic description information of the industrial robot ontology to form a structured and logically reasoning cognitive knowledge graph, including the perception layer and the cognitive layer. The realization of automatic representation of robot perception information enhances the versatility, systematicness, and intuitiveness of robot cognitive information representation and can effectively improve the cognitive reasoning ability and knowledge retrieval efficiency of robots in the industrial Internet environment.

Patent US20200317093A1 proposed a query response system for converting natural language queries into standard queries using neural networks, with a processor that determines the relevance of documents and returns documents when they are determined to be relevant. This application describes a system and method for converting natural language queries into standard queries using sequence-to-sequence neural networks. As described in this article, when a natural language query is received, the natural language query is converted into a standard query using a sequence-to-sequence model. In some cases, the sequence-to-sequence model is associated with the layer of interest. The perform searches using standard queries and can return various documents. The documents obtained by the search are scored based at least in part on the determined conditional entropy of the documents. Use natural language queries and documents to determine conditional entropy.

6.2. Deep Learning. The importance of algorithms related to AI and deep learning to chatbot is obvious. However, this kind of emerging technology is less noticeable in patent documents. Commonly used chatbots are LSTM, transformer, RNN, etc. Interestingly, the bidirectional mechanism is applied to almost all architectures. Chatbot-related articles using bidirectional architecture have appeared in large numbers since 2019, and their number accounted for more than 80% of all years (see Table 13).

Patent CN111267097A proposed a natural language-based industrial robot-assisted programming method, including parsing language instructions, matching analysis results, and combining coordinate output to generate the final robot-assisted code. The present invention requires a method for auxiliary programming of natural language-based industrial robots according to language instructions and generating corresponding executable codes for the environment image robot. The present invention is divided into three parts. First, use LSTM bidirectional recurrent neural network (Bi-RNN) and fast regional convolutional neural network (F-RCNN) to extract language instructions and features of the factory environment. Second, provide the "attention mechanism" model of the alignment algorithm, and correctly match the machine translation of the instruction in the machine environment, so as to identify the specified object and the output coordinate point of the object. Third, use the model output of the generating operation to match the CoBlox result modular programming model.

The technical development of DL in NLP has been quite mature. Although academic research is constantly pursuing better performance, it is already more than enough at the applied level. When applying any framework commonly used today, even with little training data, a chatbot is able to be perceived satisfactory by users [57]. Therefore, in addition to being used to handle NLP tasks, the other main application of DL is to assist the dialogue management of the chatbot system.

Patent CN108282587B proposes a mobile customer service dialogue management method based on state tracking and policy orientation for communication industry,

Table 13: Bidirectional related article number.

Search terms	No. of results	
	All	Since 2019
Chatbot Bi-RNN	30	25
Chatbot BiLSTM	422	347
Chatbot BERT	1,290	1,150

involves adopting the deep Q-network-based strategy optimization method to select best action strategy. The method involves establishing a dialogue problem guiding strategy based on the partially observable Markov decision process (POMDP) model, and applying an action to dialogue environment state of user through the internal action of the POMDP model, so that the state of the conversation environment changes and a certain return is obtained. The likelihood of executing a series of strategies is measured based on the cumulative returns obtained, and the problem is turned into a strategy choice problem. A deep-enhanced learning problem-guided strategy optimization algorithm is constructed based on the dialogue problem guiding strategy obtained by the POMDP model, and a deep Q-network (DQN)-based strategy optimization method is adopted to select the best action strategy.

6.3. Speech-Related Technologies. Chatbot has developed towards an integrated conversation system, where in the context of multiperson conversations, speech segmentation and speaker recognition algorithms have been the main research topics in recent years [58, 59]. Li et al. [60] summarizes the modern noise-robust technology of ASR developed in the past 30 years and proposes the classification standards for various noise-robust technologies, and the pros and cons of using different antinoise ASR technologies in actual application scenarios. For example, for stable voice-controlled driving, the environmental conditions of drones must be handled carefully, including environmental noise that can reduce the accuracy of recognition. So, Park and Na [61] studied multiple unmanned aerial vehicle (UAV) control and noise reduction methods driven by voice.

Patent CN111768768A proposes a method of processing voice in the fields of AI, DL, NLP and voice interaction, and noise reduction processing on voice data sent by peripheral control equipment. The specific implementation scheme is as follows: in response to the acquired voice recognition interface call request sent by the peripheral control device, start the voice recognition process; acquire the type of the peripheral control device; determine the target voice noise reduction mode according to the type of the peripheral control device. In the noise mode, noise reduction is performed on the voice data sent by the peripheral control device to obtain the voice data after noise reduction; after noise reduction, voice recognition is performed on the voice data to generate text data. Therefore, through the voice processing method, the noise level generated by other operations in the peripheral control device included in the voice data is reduced.

6.4. Speech-Driven Automated Control. Interactive Smart Agents (ISAs), which are controlled by users through natural language dialogues, are becoming a part of life, especially in smart home scenarios [62]. Patent WO2020203067A1 describes an information-processing device containing a control unit driven by natural language, which is arranged for controlling the movement of a moving object on the basis of results of a speech recognition process. Patent CN110654738A describes an automatic garbage classification and recycling device based on NLP. The garbage bins are, respectively, equipped with infrared sensors, and the lower box body is equipped with a mechanical transmission mechanism and an automatic classification mechanism. The device and method of the present invention have high recognition efficiency and high degree of automation.

6.5. Internet of Things (IoT). Patent KR2020131299A (assignee: Google LLC) proposes a method of associating multiple remote automation assistant components through IoT devices, combined with voice recognition modules to monitor and send voice data. Patent US10543931B2 proposes a method for monitoring audible and message alerts received during flight in the aircrafts. IoT cockpit includes subsequently marking a cascaded message alert to associate with the display element. After receiving a plurality of alerts, including at least one of the audible alerts or message alarm, the first NLP task is applied to convert the auditory alarm into a text alarm that is structurally consistent with the format for aggregation, or a cascaded message alarm, where the second NLP task is applied to identify the context.

6.6. Applied Scenarios. According to the A-TFM results in Section 5.4, it can be found that the related patents of chatbot applications are still mainly focused on personalized services and e-commerce. Both types of applications are focused on using chatbot as a virtual assistant serving a specific purpose, or using chatbot as an expert in a specific field to achieve the purpose of knowledge acquisition. These applications for providing utility or productivity are progressing towards education [63, 64], medical [65], emotional [66, 67], and social services [68–70]. Under these conditions, the integration of socioemotional behavior and personality processing design principles can lead to a decisive competitive advantage [71]. The application trend of chatbot obtained from the patent analysis in this study is consistent with some studies [71, 72], which illustrates the effectiveness of this research.

7. Conclusion

The study conducts a comprehensive patent review on emerging technologies of natural language-enabled chatbots. The contribution of this study is addressed in Section 7.1, the managerial implication is described in Section 7.2, the practical/social implications for marketers are described in Section 7.3, and the limitations and future research are suggested in Section 7.4.

7.1. Contribution. The contribution of this study is from three aspects. First, a patent analytic framework is proposed and proved to be effective. Second, emerging technologies are found. Third, application trend is addressed.

A patent analytic framework starts from patent-based ontology construction, followed by patent management map and TFM, and performing the case study part. The four-level hierarchical structure of the ontology is constructed with text-mining approaches such as k-means clustering algorithm and LDA topic modeling, to reduce human interference during the process. The ontology map can be used as the basis for strategic and sustainable R&D planning, from which researchers are able to quickly understand the development trends of key technologies and can identify technology gaps. It is worth noting that in some past patent analysis articles, detailed patent query conditions were first designed, on which the following analysis are based [25]. However, the patent analysis method proposed in this research uses iterative process to find out the most appropriate query conditions and patent information during the construction of ontology. In addition to patent analysis, it is reasonable to find emerging technologies from academic articles, and systematic literature review (SLR) is the main method. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) was created by the international health collaborator network and provides a framework for the SLR to ensure methodological rigor and quality [73]. The conduct of an SLR largely depends on the scope and quality of the included research. Therefore, the systematic reviewer may need to modify its original evaluation plan in the process, and the PRISMA statement recognizes this iterative process [74]. This provides crucial support for the iterative method used in this study to continuously adjust the patent query conditions in the ontology construction process.

The emerging technologies are summarized as follows. Knowledge is the basis of natural language-enabled chatbot, among which feature graph is a feature generation framework that has recently attracted attention. DL is the core of the main method, and most of the DL algorithms are mature. In recent years, patents have focused on the combination of various DL algorithms, by capturing their respective advantages and filling each other's shortcomings. In terms of speech technology, noise reduction is the focus of recent speech recognition technology. Sounds including voices and noise in operating equipment are obtained from the device and converted into refined text data through the integration of DL and NLP technologies. Furthermore, it is found that context is the main research subject, whether it is the exploration of the knowledge base or the logic of the algorithm. Previous research on NLP has focused on unstructured text, but in recent years, it has clearly turned to messages in dialogue. In unstructured texts, the term frequency-based method can have good results, but the message in the dialogue relies on a large number of pronouns and the continuity and relevance of the context, and the anaphora is more complicated. Even to be able to apply NLP to daily conversations, it faces a larger and broader domain and knowledge base. For this reason, the chatbots of various specific domains integrate with each other to become a more complete and powerful system. Communication technology and system integration are also very important.

As for the application trend, the increasing number of patents shows the rapid development of NLP chatbot in recent years. From the macroscopic patent trend analysis, the development trend of patents has been found. The patents related to natural language-enabled started in 2014 and developed rapidly since 2016. At first, it was mainly based on NLP and knowledge base. By 2018, speech recognition and communication technology have been developed and perfected, and then a large number of applications began to appear in 2019. These applications are concentrated in Silicon Valley's technology giants, and they have also brought significant improvements to people's lives. Natural language-enabled chatbot is widely used in the field of e-commerce, focusing on customer service and medical consulting. With the popularization of 5G network technology, more and more voice-driven applications, such as speech-driven automated control for IoT and system integration, along with immersive human-computer interaction interfaces provide better user experience. In addition to e-commerce applications, more applications in the product life cycle process have begun to be observed. The application scenarios of natural language-enabled chatbot have clearly begun to shift from e-commerce to engineering applications, such as product design, engineering assets management, smart manufacturing, and workshop management. Natural language-enabled chatbot, as an emerging smart system architecture using AI, has become a service integration solution through the integration of devices, algorithms, and network communication technologies. It is also expected to continue to impact the traditional information system architecture in the future.

7.2. Managerial Implication. At present, the application of chatbot is still focused on personal assistants and customer services, and these application scenarios are limited to a very limited field of knowledge

From the early rule-based dialogue interaction system to natural language interaction, coupled with the maturity of voice recognition technology, chatbot can provide good dialogue quality in chit-chat and single-round dialogue. The bottleneck of service provision has shifted from system development to the establishment of in-depth domain knowledge base. Many Internet service providers have been able to provide a convenient application framework for establishing chatbot as an automated customer service or personal service assistant. The success of the chatbot service depends on whether it accurately interprets users' context or intented question and possesses the knowledge base needed to fully support the context and provide accurate replies.

The limitation of chatbot's focus on a single domain has begun to be noticed, so the practice of integrating multiple domain chatbot into a chatbot advisory group has been seen

in recent patents and research. With the changes in chatbot system structure, multiple domain knowledges are integrated into a complex system. In recent years, the strategy of focusing on data-driven innovation has led to new products and business models in the emerging and developing digital markets. However, while exploring knowledge from data, user privacy is an issue that needs to be treated with caution [75, 76].

To sum up, the feature of chatbot shifts from simple information provision to complex information integration and versatile decision supports, which means the reasoning and automatic dialogue and interface controls must be addressed. Patents on the control of electronic devices for smart homes or cars also support this idea.

7.3. Practical/Social Implications for Marketers. The three main motivations of chatbot usage imply the importance of social media to the development of chatbot, the potential of chatbot, and immersive technology in the entertainment industry, and the issues of chatbot implementation [72]

As a platform for people to initiate conversations, social media has become main chatbot interface applications to the end users. The rapid integration of social media and chatbot in e-commerce sites continues to grow and evolve.

The second most important application motivation is entertainment, which is rarely addressed in patent documents. The realism of chatbot is still insufficient, but it can already provide rich and interesting interaction. In terms of industrial development process, VR is at a similar stage. The VR experience itself is very attractive, just like an exciting game, so the user experience when creating a virtual environment is far more important than the degree of realism [77]. It can also be found from the results of TFM that there are some patents located in chatbot combined with immersive technology to improve user experience. For digital marketers, it implies that combining VR and chatbot in marketing and entertainment is expected to bring users a more immersive and innovative experience.

The third most application motivation is about social services, such as social care for the elderly living alone. In the 3D-TFM proposed in this research, some patents for chatbot applications in social services and education scenarios have indeed been observed. The Turing Test was proposed in 1950 as a method to examine how a machine behaves like a person [78]. In 2000, 50 years later, there has been a lot of controversy about the relationship between the Turing Test and AI development [79]. However, now, with the mature development of DL technology nowadays that brings clear productivity and benefits, it is not that important whether a chatbot behaves like a person. An article on the application of chatbot in health care also mentioned that "AI needs to pass the implementation game, not the imitation game" [80]. The applications of service industries, such as entertainment, social service, and education, imply that chatbot should not be regarded as merely an emulated person, but a system interface that can talk in natural language and should be more convenient for human-computer interactions. Although studies have shown that consumers generally prefer to interact with people compared to chatbots, giving human qualities can still effectively enhance the consumer experience [81]. For marketers, it will be an important issue to strike a balance between competent tasks and anthropomorphic enthusiastic responses.

7.4. Limitations and Future Research. The first limitation is that the data source selected for this study is patent documents from the DI collective global database

The smart search feature of the DI database uses natural language processing and deep learning methods to help find related patents that match the user's domain description. Compared with the traditional field search, this is a great feature that can help identify related patents faster and more accurately. Nonetheless, this limits the use of paid DI database for comprehensive patent set. The second limitation is that even though data-driven ontology construction methods are investigated in this study, domain experts are still needed to be involved in the entire operation of the framework for two main purposes, key term extraction and result verification. When searching for patents in a specific domain, relevant term will appear in a large number of patent documents. Although the TD-IDF vectorization mechanism has considered both the number of terms and the uniqueness in all documents, the clustering results show that each cluster still contains a large number of common terms. In the results of topic modeling, these general terms are the main topics corresponding to the clustering results, which indirectly confirms the validity of the method of this research. However, even though we construct ontology from patent documents through a data-driven method, we still need domain experts to verify the correctness of its ontology. In addition, in the construction process of TFM, this research also explores the scenarios in which these technologies and functions are applied. Terms related to these scenarios are mentioned in patent data but occupy little number of words. This is also a limitation on TF-based text-mining method.

Future research will solve the problems mentioned above. The first is to expand the source of data. In addition to patent data, Ribeiro-Navarrete et al. [76] proposed an SLR method of analyzing academic articles or the nonpatent literature. It is expected that a more comprehensive view might be provided by adding SLR in future research, and the comparison between the results of SLR and patent-mining can be further investigated. Moreover, how to better eliminate repeated terms in unstructured documents iteratively or other approaches will help to make text-mining methods more focused on finding unique representing terms in specific domain. Thus, since quantitative and similarity-based text-mining approaches have been applied and reach the limit, advanced technologies related to key term identification are clearly very important future research. Despite the above limitations, the framework proposed in this study, which analyzes the development of natural language-enabled chatbot with quantitative supporting data, finds

emerging technologies and points out possible future development directions and is still comprehensive and effective. In addition, this method and framework are universal and can be easily applied to discover emerging technologies in other domains.

The patent analysis method proposed in this research is used to explore the emerging technologies and trends of natural language-enabled chatbot, which can reach high consistency with the hints given in academic research. The methodology of this research is not restricted by a specific domain, so the authors hope that this methodology can be used as a reference for researchers to explore more emerging technologies and trends in other fields, so as to demonstrate the contribution of this research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partially supported by research grant funded by the Ministry of Science and Technology (grant no. MOST-108-2221-E-007-075-MY3). The authors also express their gratitude to Yi-An Su for helping refine the illustrations in the paper.

References

[1] E. Ntoutsi, P. Fafalios, U. Gadiraju et al., "Bias in data-driven artificial intelligence systems—An introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[2] L. Goasduff, *2 Megatrends Dominate the Gartner Hype Cycle for Artificial Intelligence*, Gartner, Inc., Stanford, CT, USA, 2020, https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/.

[3] L. Wood, *Global Chatbot Market Anticipated to Reach $9.4 Billion by 2024 - Robust Opportunities to Arise in Retail & eCommerce*, Insider Inc., New York, NY, USA, 2019, https://markets.businessinsider.com/news/stocks/global-chatbot-market-anticipated-to-reach-9-4-billion-by-2024-robust-opportunities-to-arise-in-retail-ecommerce-1028759508.

[4] A. Chopra, *21 Vital Chatbot Statistics for 2020*, OutGrow, New York, NY, USA, 2020, https://outgrow.co/blog/vital-chatbot-statistics.

[5] R. Dale, "The return of the chatbots," *Natural Language Engineering*, vol. 22, no. 5, pp. 811–817, 2016.

[6] A. S. Gillis, *Turing Test*, TechTarget, Newton, MA, USA, 2019, https://searchenterpriseai.techtarget.com/definition/Turing-test.

[7] S. Reeves, V. Williams, F. M. Costela et al., "Narrative video scene description task discriminates between levels of cognitive impairment in Alzheimer's disease," *Neuropsychology*, vol. 34, no. 4, pp. 437–446, 2020.

[8] J. Dai and Z. Ma, "Automatic identification of bond information based on OCR and NLP," *Journal of Computers*, vol. 14, no. 6, pp. 397–403, 2019.

[9] V. K. Jain and S. Kumar, "Predictive analysis of emotions for improving customer services," in *Natural Language Processing: Concepts, Methodologies, Tools, and Applications*, pp. 808–817, IGI Global: Hershey, PA, USA, 2020.

[10] M. Baez, F. Daniel, F. Casati, and B. Benatallah, "Chatbot integration in few patterns," *IEEE Internet Computing*, vol. 99, 2020.

[11] L. Goasduff, *Chatbots Will Appeal to Modern Workers*, Gartner, Inc., Stanford, CT, USA, 2019, https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers.

[12] C.-C. Chang and C.-Y. Liu, "Using patent deployment to support industrialization of new technology," *Intellectual Property Rights Monthly*, vol. 224, pp. 6–21, 2017.

[13] Y. Keng, *Application of Patent Information to Business Planning of Enterprises*, Judicial Yuan, Taipei, Taiwan, 2019, https://www.judicial.gov.tw/tw/cp-1429-66877-ae6e7-1.html.

[14] J. R. Saura, "Using data sciences in digital marketing: framework, methods, and performance metrics," *Journal of Innovation & Knowledge*, vol. 6, 2020.

[15] B. Yoon and Y. Park, "A text-mining-based patent network: analytical tool for high-technology trend," *The Journal of High Technology Management Research*, vol. 15, no. 1, pp. 37–50, 2004.

[16] A. Abbas, L. Zhang, and S. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, 2014.

[17] Y. G. Kim, J. H. Suh, and S. C. Park, "Visualization of patent analysis for emerging technology," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1804–1812, 2008.

[18] G. Kim and J. Bae, "A novel approach to forecast promising technology through patent analysis," *Technological Forecasting and Social Change*, vol. 117, pp. 228–237, 2017.

[19] M. Thilakaratne, K. Falkner, and T. Atapattu, "A systematic review on literature-based discovery workflow," *PeerJ Computer Science*, vol. 5, p. e235, 2019.

[20] U. H. Govindarajan, A. Trappey, and C. Trappey, "Immersive technology for human-centric cyberphysical systems in complex manufacturing processes: A comprehensive overview of the global patent profile using collective intelligence," *Complexity*, vol. 2018, Article ID 4283634, 14 pages, 2018.

[21] V. Singh, K. Chakraborty, and L. Vincent, "Patent database: Their importance in prior art documentation and patent search," *Journal of Intellectual Property Rights*, vol. 21, 2016.

[22] O. Krejcar, R. Frischer, R. Hlavica, K. Kuca, P. Maresova, and A. Selamat, "Review of available SW solutions for intellectual property management systems from the perspective of open innovation," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 2, p. 23, 2020.

[23] H. Chiu, *Smart Search*, Clarivate, Philadelphia, PA, USA, 2021, https://clarivate.com.tw/blog/2017/11/08/derwent-innovation-ai-smart-search/.

[24] S.-S. Weng, H.-J. Tsai, S.-C. Liu, and C.-H. Hsu, "Ontology construction for information classification," *Expert Systems with Applications*, vol. 31, no. 1, pp. 1–12, 2006.

[25] A. J. C. Trappey, C. V. Trappey, and A.-C. Chang, "Intelligent extraction of a knowledge ontology from global patents," *International Journal on Semantic Web and Information Systems*, vol. 16, no. 4, pp. 61–80, 2020.

[26] D. Tsatsou, P. C. Davis, J. Li, I. Kompatsiaris, and S. Papadopoulos, *Ontology Construction*, Google Technology Holdings LLC, Mountain View, CA, USA.

[27] R. Subhashini and J. Akilandeswari, "A survey on ontology construction methodologies," *International Journal of Enterprise Computing and Business Systems*, vol. 1, no. 1, pp. 60–72, 2011.

[28] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, no. 5, pp. 1216–1247, 2007.

[29] G. Battsengel, S. Geetha, and J. Jeon, "Analysis of technological trends and technological portfolio of unmanned aerial vehicle," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 3, p. 48, 2020.

[30] D. Korobkin, S. Fomenkov, A. Kravets, and S. Kolesnikov, "Methods of statistical and semantic patent analysis," in *Proceedings of the Conference on Creativity in Intelligent Technologies and Data Science*, Springer, Volgograd, Russia, September 2017.

[31] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, and J. Hu, "Patent keyword extraction algorithm based on distributed representation for patent classification," *Entropy*, vol. 20, no. 2, p. 104, 2018.

[32] T. Shanie, J. Suprijadi, and Zulhanif, "Text grouping in patent analysis using adaptive K-means clustering algorithm," in *Proceedings of the AIP Conference 2017*, AIP Publishing LLC, Bikaner, India, November 2017.

[33] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, 2018.

[34] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning BERT language model," *World Patent Information*, vol. 61, Article ID 101965, 2020.

[35] S. Jun, "Technology integration and analysis using boosting and ensemble," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 7, no. 1, p. 27, 2021.

[36] *Technology Function Matrix Analysis-Quickly Grasp the Patent Layout and Explore the Blue Ocean Opportunities*, Wideband IP Office, Taipei, Taiwan, 2017, http://www.widebandip.com/tw/mobile/knowledge2.php?type1=B&idno=224.

[37] A. J. C. Trappey, C. V. Trappey, U. H. Govindarajan, and A. C. C. Jhuang, "Construction and validation of an ontology-based technology function matrix: technology mining of cyber physical system patent portfolios," *World Patent Information*, vol. 55, pp. 19–24, 2018.

[38] Y. T. Demey and D. Golzio, "Search strategies at the European patent office," *World Patent Information*, vol. 63, Article ID 101989, 2020.

[39] Y. Yang and G. Ren, "HanLP-based technology function matrix construction on Chinese process patents," *International Journal of Mobile Computing and Multimedia Communications*, vol. 11, no. 3, pp. 48–64, 2020.

[40] W. Ki and K. Kim, "Generating information relation matrix using semantic patent mining for technology planning: a case of nano-sensor," *IEEE Access*, vol. 5, pp. 26783–26797, 2017.

[41] A. J. C. Trappey, C. V. Trappey, T. M. Wang, and M. Y. L. Tang, "Ontology-based technology function matrix for patent analysis of additive manufacturing in the dental industry," *International Journal of Manufacturing Research*, vol. 12, no. 1, pp. 64–82, 2017.

[42] *Analysis of Common Reasons for Rejection of Taiwan's Artificial Intelligence-Related Patents*, IP Office, Ministry of Economic Affairs, Taipei, Taiwan, 2019, https://www.tipo.gov.tw/tw/cp-85-859330-1189b-1.html.

[43] *Everything You Need to Get Started in Medical Billing & Coding*, Medical Billing and Coding.org, a Red Ventures Company, Indian Land, SC, USA, 2021, https://www.medicalbillingandcoding.org/medical-billing-coding/.

[44] X. Han, F. Zhou, Z. Hao et al., "MAF.-CNER:A Chinese named entity recognition model based on multifeature adaptive fusion," *Complexity*, vol. 2021, Article ID 6696064, 9 pages, 2021.

[45] D. Adiwardana, M.-T. Luong, D. R. So et al., "Towards a human-like open-domain chatbot," 2020, http://arxiv.org/abs/2001.09977.

[46] S. Roller, E. Dinan, N. Goyal et al., "Recipes for building an open-domain chatbot," 2020, http://arxiv.org/abs/2004.13637.

[47] S. Bao, H. He, F. Wang et al., "Plato-2: Towards building an open-domain chatbot via curriculum learning," 2020, http://arxiv.org/abs/2006.16779.

[48] S. S. Abdullahi, S. Yiming, A. Abdullahi, and U. Aliyu, "Open domain chatbot based on attentive end-to-end Seq2Seq mechanism," in *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, Sanya, China, December 2019.

[49] C. H. Hong, Y. Liang, S. S. Roy et al., "Audrey: A personalized open-domain conversational bot," 2020, http://arxiv.org/abs/2011.05910.

[50] P. Rastogi, A. Gupta, T. Chen, and L. Mathias, "Scaling multi-domain dialogue state tracking via query reformulation," 2019, http://arxiv.org/abs/1903.05164.

[51] D. Calvaresi, J.-P. Calbimonte, E. Siboni et al., "EREBOTS: privacy-compliant agent-based platform for multi-scenario personalized health-assistant chatbots," *Electronics*, vol. 10, no. 6, p. 666, 2021.

[52] C.-Y. Li, D. Ortega, D. Väth et al., "ADVISER: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents," 2020, http://arxiv.org/abs/2005.01777.

[53] N. A. Ahmad, M. H. Che, A. Zainal, M. F. Abd Rauf, and Z. Adnan, "Review of chatbots design techniques," *International Journal of Computer Applications*, vol. 181, no. 8, pp. 7–10, 2018.

[54] A. Ait-Mlouk and L. Jiang, "KBot: a Knowledge graph based chatBot for natural language understanding over linked data," *IEEE Access*, vol. 8, pp. 149220–149230, 2020.

[55] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, *Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371)*Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Wadern, Germany, 2019.

[56] D. Zhong, Y.-A. Zhu, L. Wang, J. Duan, and J. He, "A cognition knowledge representation model based on multi-dimensional heterogeneous data," *Complexity*, vol. 2020, Article ID 8812459, 17 pages, 2020.

[57] N. Tavanapour and E. A. Bittner, "Automated Facilitation for Idea Platforms: Design and Evaluation of a Chatbot Prototype," in *Proceedings of the Thirty Ninth International Conference on Information Systems*, San Francisco, CA, USA, 2018.

[58] Z. Su, Y. Li, and G. Yang, "Dietary composition perception algorithm using social robot audition for Mandarin Chinese," *IEEE Access*, vol. 8, pp. 8768–8782, 2020.

[59] Q. Li, R. Gravina, Y. Li, S. H. Alsamhi, F. Sun, and G. Fortino, "Multi-user activity recognition: challenges and opportunities," *Information Fusion*, vol. 63, pp. 121–135, 2020.

[60] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[61] J.-S. Park and H.-J. Na, "Front-end of vehicle-embedded speech recognition for voice-driven multi-UAVs control," *Applied Sciences*, vol. 10, no. 19, p. 6876, 2020.

[62] G. Spinelli, U. Shamim, A. Woodcock, and A. Nair, "Enabling by Voice. Voice Enabled Environmental Control (EC) Devices Using Interactive Smart Agents," in *Proceedings of the Twenty-*

fifth Americas Conference on Information Systems, ISAs, Cancun, Mexico, 2019.

[63] G. Vladova, J. Haase, L. S. Rüdian, and N. Pinkwart, *Educational Chatbot with Learning Avatar for Personalization*, 2019.

[64] S. Gupta, K. Jagannath, N. Aggarwal et al., "Artificially Intelligent (AI) Tutors in the Classroom: A Need Assessment Study of Designing Chatbots to Support Student Learning," in *Proceedings of the PACIS 2019*, X'ian, China, 2019.

[65] F. Mehfooz, S. Jha, S. Singh, S. Saini, and N. Sharma, "Medical chatbot for novel COVID-19," in *ICT Analysis and Applications*, pp. 423–430, Springer, Berlin, Germany, 2021.

[66] F. Catania, N. Di Nardo, F. Garzotto, and D. Occhiuto, "Emoty: an emotionally sensitive conversational agent for people with neurodevelopmental disorders," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, Maui, HI, USA, January 2019.

[67] M. Jaiswal, C.-P. Bara, Y. Luo et al., "Muse: a multimodal dataset of stressed emotion," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020.

[68] A. Augello, M. Gentile, L. Weideveld, and F. Dignum, "A model of a social chatbot," in *Smart Innovation, Systems and Technologies*, vol. 55, pp. 637–647, Springer, 2016.

[69] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, CO, USA, May 2017.

[70] T. Hu, A. Xu, Z. Liu et al., "Touch your heart: a tone-aware chatbot for customer care on social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, QC, Canada, April 2018.

[71] A. Janssen, J. Passlick, D. Cordona, and M. Breitner, "Virtual assistance in any context: a taxonomy of desgin elements for domain-specific chatbots," *Business & Information Systems Engineering*, vol. 62, 2020.

[72] P. B. Brandtzaeg and A. Følstad, "Why people use chatbots," in *Proceedings of the International Conference on Internet Science*, Springer, Florence, Italy, November 2017.

[73] D. Pati and L. N. Lorusso, "How to write a systematic review of the literature," *HERD: Health Environments Research & Design Journal*, vol. 11, no. 1, pp. 15–30, 2018.

[74] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and a. t. P. Group, "Reprint-preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *Physical Therapy*, vol. 89, no. 9, pp. 873–880, 2009.

[75] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marqués, "From user-generated data to data-driven innovation: a research agenda to understand user privacy in digital markets," *International Journal of Information Management*, vol. 2, Article ID 102331, 2021.

[76] S. Ribeiro-Navarrete, J. R. Saura, and D. Palacios-Marqués, "Towards a new era of mass data collection: assessing pandemic surveillance technologies to preserve user privacy," *Technological Forecasting and Social Change*, vol. 167, p. 120681, 2021.

[77] M. S. van Gisbergen, I. Sensagir, and J. Relouw, "How real do you see yourself in VR? The effect of user-avatar resemblance on virtual reality experiences and behaviour," in *Augmented Reality and Virtual Reality*, pp. 401–409, Springer, Berlin, Germany, 2020.

[78] A. M. Turing, "I.-Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[79] A. P. Saygin, I. Cicekli, and V. Akman, "Turing test: 50 years later," *Minds and Machines*, vol. 10, no. 4, pp. 463–518, 2000.

[80] J. Powell, "Trust me, I'ma chatbot: how artificial intelligence in health care fails the Turing test," *Journal of Medical Internet Research*, vol. 21, no. 10, Article ID e16222, 2019.

[81] R. Roy and V. Naidoo, "Enhancing chatbot effectiveness: the role of anthropomorphic conversational styles and time orientation," *Journal of Business Research*, vol. 126, pp. 23–34, 2021.

Hindawi
Complexity
Volume 2021, Article ID 5583287, 14 pages
https://doi.org/10.1155/2021/5583287

Research Article

MyOcrTool: Visualization System for Generating Associative Images of Chinese Characters in Smart Devices

Laxmisha Rai and Hong Li

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Correspondence should be addressed to Laxmisha Rai; laxmisha@ieee.org

Received 6 February 2021; Revised 12 March 2021; Accepted 16 April 2021; Published 11 May 2021

Academic Editor: Abd E.I.-Baset Hassanien

Majority of Chinese characters are pictographic characters with strong associative ability and when a character appears for Chinese readers, they usually associate with the objects, or actions related to the character immediately. Having this background, we propose a system to visualize the simplified Chinese characters, so that developing any skills of either reading or writing Chinese characters is not necessary. Considering the extensive use and application of mobile devices, automatic identification of Chinese characters and display of associative images are made possible in smart devices to facilitate quick overview of a Chinese text. This work is of practical significance considering the research and development of real-time Chinese text recognition, display of associative images and for such users who would like to visualize the text with only images. The proposed Chinese character recognition system and visualization tool is named as MyOcrTool and developed for Android platform. The application recognizes the Chinese characters through OCR engine, and uses the internal voice playback interface to realize the audio functions and display the visual images of Chinese characters in real-time.

1. Introduction

In recent years, increasing number of foreigners are studying Chinese as a second language around the world through institutions that promote Chinese language education. As many foreigners are coming to China, either to study, work, and travel—reading, speaking, listening and writing Chinese language are one of the essential requirements for managing the daily activities. Comparing to several other major foreign languages, Chinese languages is considered as one of the hardest language for learners [1]. To read a newspaper or an online article, one need to master at least over two thousand characters. As there are nearly several thousands of Characters in Chinese language, it is essential for a learner to master at least around two to three thousand characters to read or understand a sign board, public safety instructions or a food menu. In [2], a website lists nearly 4000 simplified Chinese characters, based on their frequency of their appearance in written documents. As per the website, a solid knowledge of all these characters makes a learner to read any document written in simplified Chinese. However, for travelers who are interested in short-term stay in China, they may be more interested in understanding the meaning behind the text, rather than spending countless hours on studying Chinese or hiring a translator. Moreover, these days several software based applications, and mobile Apps are popular, which help the learners to understand the meaning of a Chinese character or a word. In the past years, there are several modern approaches to simplify learning Chinese using mobile devices, software tools, Apps, and electronic devices. A study on mobile assisted language learning (MALL) that emphasizes learner created content and contextualized creation of meanings is presented in [3]. While learning Chinese idioms, the tool let the students to take initiatives to use mobile phones to capture scenes that express the meaning of idioms, and to facilitate to construct sentences with them. This way of transforming idioms into photos can help students understand idioms in a more efficient way. You and Xu [4] evaluated the usability of a system named Xi-Zi-e-Bi-Tong (習-e-筆通), which is one of

the systems for writing Chinese characters used by the education ministry. The main focus of this study is to evaluate the efficacy of the system for foreign learners from different cultural backgrounds and ages. In summary, although Chinese non-native speakers can interact with the system, there are still problems exist, and there is scope for improvement. In [5], researchers designed and evaluated a software that facilitate users to learn Chinese through the use of mobile application. The results from the past literature shows that vast majority of foreign learners are satisfied with learning Chinese through electronic devices, and this also plays a huge role in learning Chinese.

To use the online or mobile applications, and language dictionaries related to Chinese language, a user need to aware of three things. (1) user should be able to know how to read a Chinese character, (2) user need to know how to write a Chinese character, mostly on phone screen by drawing different strokes and following stroke order and, (3) user need to be aware of usage of pinyin (*pīnyīn*). However, as mentioned earlier, as there are thousands of Chinese characters, and Chinese is a complex language, it is not easy for a non-native learner to be aware of all these.

Majority of Chinese characters represent some actions, events, animals, humans, or objects directly or indirectly. It is evident from Figure 1, that, the Chinese characters are evolved by following logical rules over the years. There are several stages of evolution of Chinese characters [6]. Some of these stages include oracle bone script, bronze script, small seal script, clerical script, standard script and simplified Chinese as shown in Figure 1. Today, simplified Chinese is the most common and widely used script for all official purposes in China. By merely grasping a Chinese character, its connotation and extension, it can produce endless associations. So, in theory, for a Chinese learner in the early stages, commonly used characters are presented as a painting or picture to provide quick association with that character. Earlier, a comprehensive analysis of the images generated by the simplified Chinese characters through the use of Internet and electronic devices, and emphasizing the understanding of text in the form of limited images is studied in [7]. This work is inspired by majority of opinions of scholars, where no matter how simple or complicated a character is, the Chinese character is still a picture. Here, the researchers, tried to understand how each character is represented as images in Internet usage, and in popular messaging tools. With this background, we try to investigate answers to the following research questions. (1) Considering our previous study [7], is it possible to visualize simplified Chinese characters in real-time with their associative images using smart devices? (2) How to develop the application for real-time visualization of Chinese characters extracted from different sources facilitating the evaluation of recognition rates?

The first research question is related to development of an visualization system for generating associative images of Chinese characters. Considering this aspect, there are several related works, which describe the research related to visual perception, virtual reality in applications of industrial domains, and Internet of Things (IoT) in the recent years. In [8] authors have provided detailed account of applications of visual perceptions in different industrial fields. Thee industrial fields include agriculture, manufacturing, and autonomous driving. In [9], importance of human visual system while acquiring different features of an image, and the impact of distortion distribution within an image is studied. In [10], a metric for evaluation of screen contents images for better visual quality is explored. In [11], researchers constructed an image processing and quality evaluation system using convolutional neural network (CNN), and IoT technology to investigate the applications of industrial visual perception in smart cities. The main goal is to provide experimental framework for future smart city visualization. In [12], as an security solution framework, an intrusion detection model of industrial control network is designed and simulated in virtual reality (VR) environment. In [13], the relevance of VR technology applications with consideration to IoT is discussed.

Chinese characters are pictographic characters (or pictograms) with strong associative ability and when a character appears for Chinese readers, they usually associate with the objects, or actions related to the character immediately. Having this background, we propose a system to visualize the simplified Chinese characters so that non-native learners can understand the meaning of a character quickly without even typing or learning it. Considering the extensive use and application of mobile devices, automatic identification of Chinese characters and display of associative images are made possible in smart phones to facilitate quick overview of a Chinese text. This work is of practical significance considering the research and development of real-time Chinese text recognition and display of associative images for such users who has no background in Chinese writing or reading. The proposed Chinese character recognition system and visualization tool is named as MyOcrTool and is suitable for Android platform. The application recognizes the Chinese characters through optical character recognition (OCR) engine called Tesseract, and uses the internal voice playback interface to realize audio functions for character pronunciation and display the visual images of Chinese characters in real-time.

The main purpose of this study is to generate images for each Chinese character and then a representative image for the entire text, so that a user can able to obtain the approximate idea presented in the text. Moreover, a user is able to obtain the meaning, even without understanding *pinyin*, or romanization or any reading ability. So, the system is useful for anyone, who has no literacy on Chinese language, or someone who is unable to listen or speak.

Table 1 shows a list of examples of selected characters and their associated images. This kind of visual images for Chinese characters able to help non-Chinese to visualize the meaning behind Chinese characters rapidly. The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes model description by providing details of OCR, Tesseract open source engine, application overview, system design and implementation details. Section 4 presents details of experimental design and analysis of results. Finally, we conclude the paper in Section 5 with some pointers to future work.

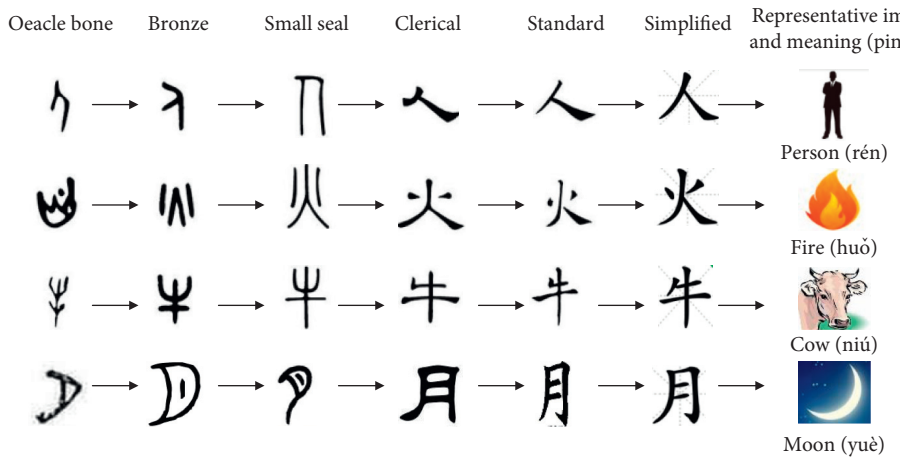


Figure 1: Some stages of evolution of Chinese characters.

Table 1: Sample list of examples of characters and their associated images.

Characters	Meaning	Pinyin	Associative images
心	Heart	xīn	
球	Ball	qiú	
书	Book	shū	
日	Sun	rì	
树	Tree	shù	
火	Fire	huǒ	
车	Car	chē	
手	Hand	shǒu	

2. Related Work

The major process required to visualize a Chinese character as image is to scan the character using smart phone camera to extract the character within a text. After scanning the character, the subsequent steps such as character recognition, display of associative image, pinyin pronunciation are performed. There are several previous related studies on character recognition in different scenarios, where studies related to characters of languages other than Chinese are also involved [14–21]. From this, one can classify all the existing text extraction methods into three categories as region based, texture based and hybrid method [14]. There are several works on text detection mentioned here, and they have proposed idea of text detection considering different factors and designed a suitable model. Kastelan et al. [15] presented the system for text extraction on image taken by grabbing the content of TV screen. An open source algorithm for OCR is used to read the text regions. After reading the text regions, comparison with the expected text is performed to make a final success or failure decision for test cases. The system successfully read the text from TV screen and used in a functional verification system. Considering the Chinese character recognition, in the past, several researchers focused their attention on recognition of printed Chinese characters [22, 23], handwritten Chinese characters [24, 25], characters in the vehicle's license plates [26, 27], and recognition of Chinese characters written in calligraphic styles [28].

In the recent years, the OCR technology is used in various applications, where recognitions of characters is the central requirement, such as in applications of e-commerce [29], and IoT [30], Moreover, the importance of Tesseract engine in character retrieval from images, translation applications, and character recognition applications is widely popular [31, 32]. Ramiah et al. [16] developed an Android application by integrating Tesseract OCR engine, Bing translator and phones' built-in speech technology. By using this App, travelers who visit a foreign country able to understand messages described in different language. Chavre and Ghotkar [17] designed an Android application, which is a user-friendly application to assist the tourist navigation, while they are roaming in foreign countries. This application is able to extract text from an image, which is captured by an mobile phone camera. The extraction is performed using stroke width transform (SWT) approach and connected component analysis. The SWT a technique used to detect texts from natural images by eliminating the noise but preserving the text. Kongtaln et al. [18] presented a method for reading medical documents by using an Android smartphone that used techniques based on the Tesseract OCR Engine to extract the text contents from medical document images such as a physical examination report. The

following factors related to the document are considered: character font, text block size, and distance between the document and the camera on the phone. Dhakal and Rahnemoonfar [19] developed a mobile application for Android platform that allows a user to take a picture of the YSI Sonde monitor (an instrument used to measure water quality parameters such as pH, temperature, salinity, and dissolved oxygen), extract text from the image and store it in a file on the phone.

Nurzam and Luthfi [20] implemented Latin text translation from Bahasa Indonesia into Javanese text with Google Mobile Vision in real-time, also vice versa with Android mobile based application. The execution flow of this design is to first scan the text through the camera, then the recognized text is transmitted to the web services. Finally, the translated text is displayed in real-time on the mobile phone screen. This research uses Javanese language or Indonesian as the outcome of the language conversion. The purpose of this research is to design and implement real-time text-based translator application using Android mobile vision which includes a combination of mobile translator application architecture and web services applications. Yi and Tian [21] proposed a method of scene text recognition from detected text regions. They first designed a discriminative character descriptor by combining several advanced feature detectors and descriptors, and then shaped the structure of the character in each character class by designing the stroke configuration map. Android system is developed to show the effectiveness of their proposed method in extracting textual information from the scene. The results of the evaluation of test data shows that their proposed text recognition scheme has a positive recognition effect, which is comparable to the major existing methods.

Moreover, other than the character recognition, there are several work which are focused on the conversion from text-to-speech (TTS). Celaschi et al. [33] integrated a set of image capturing and processing framework, such as OCR and TTS synthesis. Their work include integration of selected components and several control functions of the application: CPU through the camera to capture images; image preprocessing; OCR framework for text recognition; finally, the speech synthesis process is performed for Portuguese rather than Chinese. This design includes two versions, a preliminary desktop version designed under the Windows operating system, and a mobile device version developed as an application for Android devices. Chomchalerm et al. [34] designed an Android based App called Braille Dict that runs on smart phones. This application was developed for the blind, converting Braille input into English letters and translating them into Thai, and displaying a list of words related to the input words by retrieving them from dictionary database. One of the most significant function of this system is that, the program uses TTS function to output Thai as speech, which provides a more comfortable way for the blind to use the dictionary. In addition, several works in the past focused on OCR in android applications [35, 36], real-time OCR [37], character readability on smart phones [38], character recognition models suitable for handheld devices [39], and App to recognize food items in Chinese menu [40]. Considering several related work, it is evident that, none of the previous research focused on developing a method to visually understand the text only by scanning. So, in this paper we propose a novel method to facilitate the users to visualize the Chinese text with only by scanning it, rather than typing or entering the text into the electronic devices. Summary of these existing studies is shown in Table 2. As shown in Table 2, most of the research focuses on the OCR technology that only recognizes characters. Only three studies include text-to-speech functions. None of the studies proposes to display the visual images of Chinese characters in real-time. Therefore, this application still has its unique and innovative compared to the studies listed in the table above.

3. Model Description

3.1. OCR Technology. Considering the related works mentioned earlier, most of the earlier implementations are focused on such languages where there are limited characters in a language. However, the character extraction, and recognition is challenging especially in Chinese language considering thousands of complex Chinese characters. In this section, the problems associated with the extraction and recognition of text within an image in various scenarios is considered. Therefore, the OCR method is studied, where rapid extraction of text information from images is possible. The basic operating principle of OCR technology is to convert the information presented in documents into an image file of black and white dot matrix using camera, scanner and other optical equipments. After this process, the characters within the image are converted into editable text through the OCR engine for further information processing [41]. In recent years, OCR technology has been a hot research topic in several disciplines. The concept of OCR was first proposed by Austrian scientist Gustav Tauschek in 1929. Later, American scientist Paul Handel also proposed the idea of using technology to identify words. The earliest research on the recognition of printed Chinese characters was proposed by Casey and Nagy in 1966, where they worked on Chinese character recognition, which used template matching to identify 1000 printed Chinese characters [42]. Research work on OCR technology in China started much later. In 1970s, research on the identification of numbers, English letters and symbols began. In late 1970s, research on Chinese character recognition has started. By 1986, the study of Chinese character recognition entered a substantive stage and many research centers have successively launched Chinese OCR products. Early OCR software failed to meet actual requirements due to various factors such as recognition rate and building them as actual products.

Simultaneously, products have not reached to a level to use in practical applications due to poor execution speed and expensive hardware equipments. After 1986, China's OCR research has made substantial progress and there are several innovations on Chinese character modelling and recognition methods. The developed applications have displayed fruitful results and many centers successively launched Chinese OCR products.

Table 2: Comparison of existing character visualization systems on mobile devices.

Authors, Publication year	Mobile OCR Engine	OS	Display feature of visual images	Text-to-Speech feature
Kastelan et al., 2012	Not stated	Not stated	—	—
Kongtaln et al., 2014	Tesseract	Android	—	—
Yi and Tian, 2014	Not stated	Android	—	—
Chomchalerm et al., 2014	Not stated	Android	—	Yes
Ramiah et al., 2015	Tesseract	Android	—	Yes
Dhakal and Rahnemoonfar, 2015	Not stated	Android	—	—
Chavre and Ghotkar, 2016	Tesseract	Android	—	—
Celaschi et al., 2017	Not stated	Android	—	Yes
Nurzam and Luthfi, 2018	Not stated	Android	—	—

3.2. Tesseract Open Source Engine. The OCR technology used in this work is based on the Tesseract open source engine, which was originally developed by Hewlett-Packard (HP) between 1985 and 1994, and additional changes were made in 1996 to make it compatible with Windows [43]. In 2005, HP made available the Tesseract as open source software. Since 2006, it is developed by Google. The Tesseract engine is powerful and can be broadly divided into two parts: (1) picture layout analysis, and (2) character segmentation and recognition.

The design goal of Tesseract is character segmentation and recognition. Smith et al. [44] described the efforts to adapt the Tesseract open source OCR engine for multiple scripts and languages in 2009. They also presented the top-level block diagram of Tesseract. Real-time display of visual images associated with Chinese character is accomplished using the Android's *RecyclerView* control [45]. When the characters are recognized, it displays the visual images of respective character. In addition, the voice broadcast function use the Android's built-in TTS control [46], which does not require permission to read text and do not require Internet connection. This feature can facilitate the specified text to read aloud providing voice broadcast option to users.

3.3. Overview of the Proposed System. To answer the first research question, designing a mobile intelligent system based on platform such as Android is essential. The main function of this system is to recognize the text contained in a scanned image and display the associated image of Chinese character in real-time, and provide options for other features such as audio for character pronunciation, and pinyin display. Figure 2 shows the screenshots displaying the MyOcrTool in practical scenarios where a user using it to visualize a Chinese text. The Figure 2(a) shows a scenario, where the user trying to visualize the Chinese text in a public sign board. The Figure 2(b) shows another scenario, where the user try to visualize the restaurant menu. To operate the tool developed, a user has to follow the following steps:

(1) Open MyOcrTool and select the recognized language (Chinese, or English).

(2) Open the smart phone camera and point the scan frame to text area to be recognized.

(3) Identify the selected text area. The OCR will automatically identify the scanned text and extract valid string information.

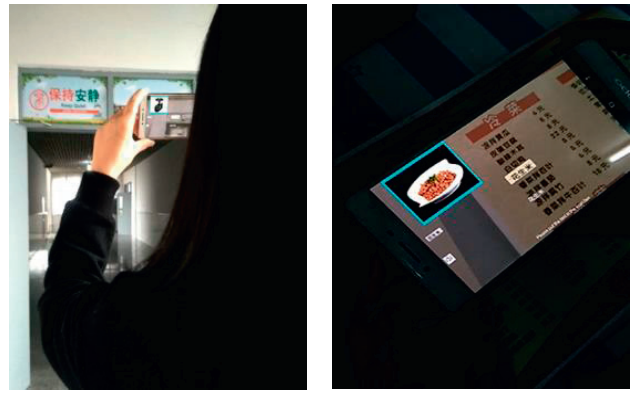
(4) Real-time display of associated pictures of Chinese characters is performed after the above steps. When a word is recognized, the associated image is displayed in real-time on the mobile phone interface.

(5) Use the voice playback feature to listen to the text recognized.

3.4. System Design and Implementation. This section introduces the overview of the system architecture and implementation details. The sequence of steps are divided into several processes. They are: (1) scanning using camera to obtain image, (2) image graying, (3) text region binarization, (4) text recognition, (5) displaying of visual images in real-time, and (6) implementation of voice broadcast feature.

3.4.1. Scanning to Obtain Images. Zxing is a Google open source library based on various 1D/2D barcode processing. It is powerful for bar code scanning and decoding via a mobile phone camera and is now commonly used to scan and decode QR codes or barcodes [47, 48]. In this work, Zxing is used to customize the scanning interface of MyOcrTool. The customization process is a three step process, which include: (1) adding the Zxing dependency packages to project, (2) configuring the permission to use the camera in manifest file, and (3) setting the scan interface and scan box.

3.4.2. Image Graying. In order for the open source engine Tesseract to better recognize the image text, some preliminary processing is needed for the image. Gray-scale is the most basic and commonly used to perform this step [49]. In the RGB model, if the values of R (red), G (green) and B (blue) are equal, then color represents a grayscale color. Moreover, the value is called grayscale value. Therefore, each pixel of grayscale image only needs one byte to store the grayscale value (also known as intensity value and brightness value), and the grayscale range is 0–255. There are four methods to gray color images: component method, maximum method, average method and weighted average method [50]. In this paper, the weighted average method is



(a) (b)

Figure 2: The screenshots displaying the MyOcrTool in practical scenarios where a user using it to visualize a Chinese text (a) in a public sign board, and (b) in a restaurant menu.

used to gray the image to obtain the image Y, and the formula is shown in equation (1). The sequence of steps and implementation details involved in picture graying is shown in Program Listing 1.

$$Y = 0.3R + 0.59G + 0.11B. \quad (1)$$

3.4.3. Text Region Binarization. In order to facilitate recognizing the text within images, binary processing of grayscale images is required [51]. Binary processing is mainly applied for the convenience of image information extraction and this can increase the recognition efficiency. Binary image refers to an image whose pixel is either black or white, and whose gray value has no intermediate transition. The most commonly used method for binarization of images is to set a threshold value T, which is used to divide the image data into two parts. The pixel groups greater than T, and groups smaller than T which are represented by 1 and 0 respectively. Considering the input grayscale image function be $f(x, y)$ and the output binary image $g(x, y)$ can be expressed by .

$$g_{(x,y)} = \begin{cases} 0, & f_{(x,y)} < T, \\ 255, & f_{(x,y)} \geq T. \end{cases} \quad (2)$$

Threshold is a measurement to distinguish the target from the background. Selecting an appropriate threshold is not only necessary to save image information as much as possible, but also to minimize the interference of background and noise, which is the principle of behind threshold selection. To accomplish this, the program use the iterative method to find the threshold [52], and this iterative method is a global binarization method. It requires the image segmentation threshold algorithm based on the approximation strategy. Firstly, an approximate threshold is selected as the initial value of the estimated value, then segmentation is performed to generate a sub-image. Following this, a new threshold is selected according to the characteristics of the sub-image and, a new threshold is utilized. Secondly, the image is divided, after several iterations, minimizing the number of incorrectly segmented image pixels. This procedure performs better than the effect of directly segmenting the image with the initial threshold. The specific algorithmic steps are as follows:

(1) Find the minimum and the maximum gray value in the image which are denoted as $Z_{\min}$ and $Z_{\max}$ respectively, then obtain the initial value of the threshold.

$$T_0 = Z_{\min} + Z_{\max}. \quad (3)$$

(2) According to threshold value T_k, the image is divided into two parts, target and background, and average gray values Z_0 and Z_1 of the two parts are obtained.

(3) Find the new threshold T_1

$$T_1 = \frac{(Z_0 + Z_1)}{2}. \quad (4)$$

(4) If $T_0 = T_1$ then the current T is the optimal threshold, otherwise the value of T_1 is assigned to T_0, and the calculation restarts from step (2).

The implementation details of the iterative method for calculating the threshold is shown in Program Listing 2.

3.4.4. Chinese Text Recognition. After the image has been pre-processed, the processed image will be used for character recognition and open source engine Tesseract is used as a tool for recognizing characters. Android Studio is used for writing programs, and programming requires Tesseract's third-party JAR package as additional support. In addition, the language package "*<language>.traineddata*" required to be placed in the mobile phone's secure digital (SD) card root directory [53]. The language packs can be downloaded directly from the Tesseract website, or its own trained language packs. This design also uses trained language packs and uses its own language library which are suitable for identifying at correct rate and speed. The flow diagram representing principle involved in character recognition is shown in Figure 3.

```
Input: original image
Output: grayImage private static Bitmap getGrayImg(){
   int alpha = 0xFF << 24;
   // Set transparency
   for (int i = 0; i < imgHeight; i++) {
      for (int j = 0; j < imgWidth; j++) {
         int grey = imgPixels[imgWidth*i + j];
         // Get the jth pixel of the i-th row
         int red = ((grey & 0x00FF0000) >> 16);
         // Get red gray value
         int green = ((grey & 0x0000FF00) >> 8);
         // Get green gray value
         int blue = (grey & 0x000000FF);
         // Get bule gray value
         grey = (int) ((float) red*0.3 + (float) green*0.59 + (float) blue*0.11);
         // obtain grayscale color values
         grey = alpha | (grey << 16) | (grey << 8) | grey; imgPixels[imgWidth*i + j] = grey;
      }
   }
   Bitmap result = Bitmap.createBitmap(imgWidth, imgHeight, Config.RGB_565);
   result.setPixels(imgPixels, 0, imgWidth, 0, 0, imgWidth, imgHeight);
   return result;
}
```

Program Listing 1: Implementation of picture graying procedure.

```
Input: grayImage
Output: threshold
private static int getIterationHresholdValue (int minGrayValue, int maxGrayValue) {
   int T1;
   int T2 = (maxGrayValue + minGrayValue)/2;
   do {
      T1 = T2;
      double s = 0, l = 0, cs = 0, cl = 0;
      for (int i = 0; i < imgHeight; i++) {
         for (int j = 0; j < imgWidth; j++) {
            int gray = imgPixels[imgWidth * i + j];
               if (gray < T1) {
                  s += gray;
                  cs++;
               }
               if (gray > T1) {
                  l += gray;
                  cl++;
               }
            }
         }
         T2 = (int) (s / cs + l / cl) / 2;
      }
      while (T1! = T2);
      return T1;
   }
```

Program Listing 2: The program listing of iterative method for calculating the threshold.

3.4.5. Real-Time Display of Associative Images. The function of displaying visual images in real-time is performed using Android's own control *RecyclerView*. *RecyclerView* is a container for displaying huge data sets that displays large volume of data in a limited window and simplifies the presentation and processing of data [45]. While using *RecyclerView*, we must specify an Adapter and a *Layout-Manager*. The main function of the Adapter is to bind the

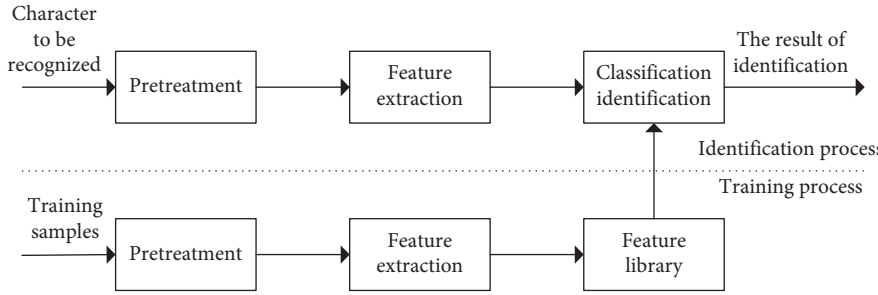


Figure 3: Flow diagram representing the principle involved in character recognition.

data to the control. The *LayoutManager* can control the layout of the *Item*. The functions of real-time display of visual images introduced in this paper are mainly to bind Chinese characters, visual pictures and edit boxes that display recognized Chinese characters. When Chinese characters are identified and displayed in the edit box, simultaneously visual images of respective characters are displayed on the mobile phone screen.

3.4.6. Implementation of Pronunciation Playback Feature. The voice playback function presented in this paper use the TTS engine that comes with Android, and it is new and important function in Android 1.6. It can be easily embedded into the application to convert the specified text into different language audio output to enhance the user experience. The role of this implementation is to play the recognized words of the Chinese text by clicking the voice button, so that the user not only able to understand the meaning of Chinese characters, but also could hear the pronunciation. The implementation details for the voice playback function is shown in Program Listing 3.

4. Results and Discussion

4.1. Overview of the Experiment. The entire application is tested on two brands of Android based mobile phones. After initially selecting the tool, we have to select the recognition language type, open the camera for scanning, and then align the scan box with the text area to be scanned. The scan frame size set by the system is of minimum width of 200 dp, a maximum width of 250 dp and a height of 80 dp, which is physically 1.5 cm wide and 0.5 cm high. The main purpose of using dp (device-independent pixel) unit is to adapt the UI layout of application to display devices of various resolutions. Finally, the recognition results and the image will appear on the mobile phone display interface as shown in Table 3, which sufficiently provide the answer to the first research question posed. The testing is carried out by considering the different font size, distance between the phone camera and text, and the text from different sources such as books, warning signs and restaurant menus. This kind of abilities to evaluate recognition rate of characters extracted from different sources answers the second research question posed in this work.

As shown in Figure 4, we have shown three general cases for displaying the images for characters and words. In case (a), where a signboard related to water conservation is translated as an image, and in case (b) for the Chinese word "中国" in a book, shows the map of China, because the word "中国" means the name of country China. In the case (c), for the word "花生米", the picture of peanut dish is displayed, because meaning of "花生米" in Chinese is peanut. We have tested for these three scenarios, with general presumption that non-native leaerners interact more frequently with signboards, restaurant menus, and tourist guidebooks.

4.2. Testing for Recognition Stability. While testing the system, we have considered the several factors as the main test criteria to evaluate the stability and recognition rate of the system. The recognition rate is defined as the ratio between the number of successfully recognized characters and the total number of characters in the test image. In Table 3, we have presented the results obtained for different kinds characters which represent animals, objects, and actions by taking sixty characters as test samples. All these characters generated independent single and unambiguous image to represent the characters and these results are considered 100% acceptable. Two reasons can be identified for this success. Traditionally these characters represent the same meaning of objects, animals and actions. Moreover, even though they are used in different contexts, and communication scenarios today, the original meaning of characters can be still possible to interpret with traditional meaning.

However, as mentioned in [6, 7], it is nearly impossible to find an exact image for each Chinese character, especially within a word because of contextual differences and usage. For example, considering an example of object "tree", some users may expect tree of bigger size, other may think tree with only few leaves, and of smaller size. So, the fundamental approach here is to provide the image which is widely accepted by users. We have also followed the similar steps as presented in [7] to collect images for testing. In some cases, several characters in a word has the same meaning, so a single image is enough to represent a word or several characters. Table 4 shows an example with several Chinese characters, their corresponding *pinyins*, and general meaning of these characters. As shown, 12 characters (如,何,吗,因,由,认,谁,思,怎, 想,若,难) may share a single image because they all have similar meanings (if, why, question?, reason, because of, how?, to recognize, difficult, who?, to think etc.).

```
 Input: text
Output: speech
private static ImageButton yuyinButton;
private TextToSpeech textToSpeech;
@Override
   protected void onCreate (Bundle savedInstanceState) {
      super.onCreate (savedInstanceState);
      setContentView (R.layout.my_scan);
      yuyinButton = (ImageButton) findViewById (R.id.yuyinButton);
      textToSpeech = new TextToSpeech (this, new TextToSpeech.OnInitListener() {
@Override
   public void onInit(int status) {
      if (status == textToSpeech.SUCCESS) {
         int result = textToSpeech.setLanguage(Locale.CHINA);
            if (result! = TextToSpeech.LANG_COUNTRY_AVAILABLE&& result !=
               TextToSpeech.LANG_AVAILABLE){
               Toast.LENGTH_SHORT).show();
            }
         }
      }
   });
yuyinButton.setOnClickListener (new View.OnClickListener() {
@Override
   public void onClick (View view) {
      textToSpeech.speak (status_view_tv_result.getText().toString(),
      TextToSpeech.QUEUE_ADD, null);
      textToSpeech.setSpeechRate (0.5f);
      textToSpeech.setPitch (0.1f);
      }
   });
}
```

Program Listing 3: The program listing of voice playback function.

Table 3: Chinese characters and generated images with 100% acceptance rate.

Objects	球	日	月	火	树	星	锁	船	车	药
	伞	杯	包	云	花	耳	鞋	脚	眼	鼻
Animals	狗	虎	兔	蛇	羊	马	牛	猴	猪	鸡
	鼠	龙	鸟	狼	象	鹰	熊	龟	猫	鱼
Actions	走	跑	看	打	唱	敲	想	听	踩	按
	喝	咳	笑	哭	读	切	吃	抱	坐	睡

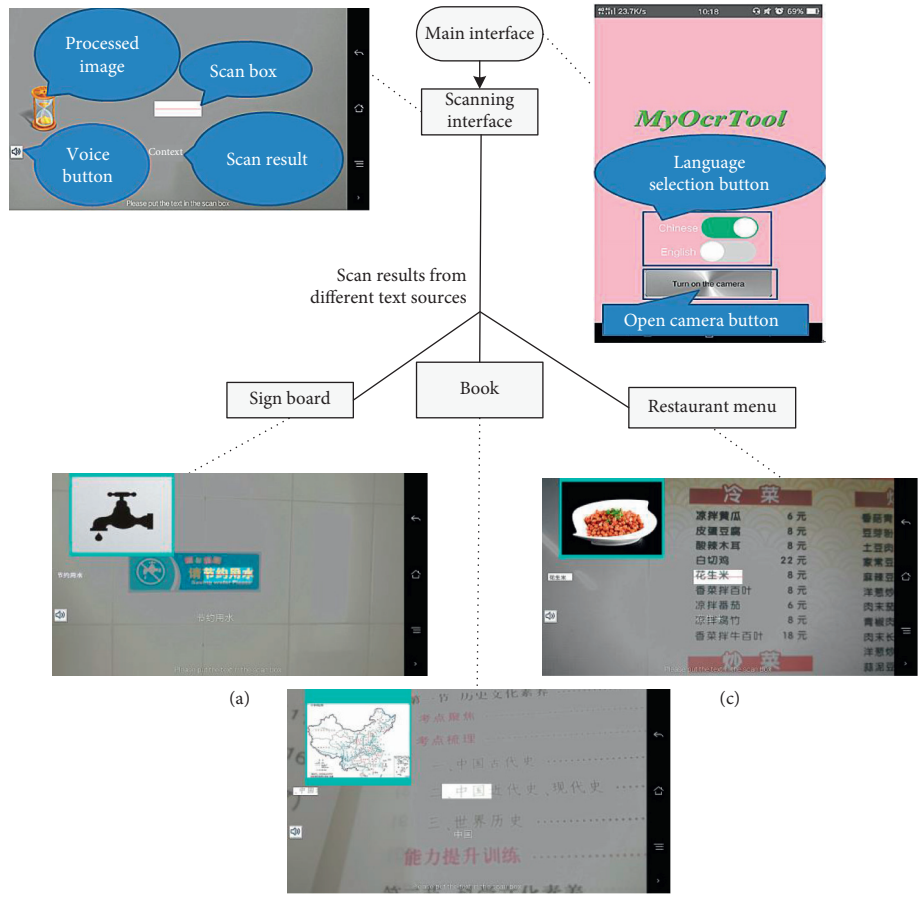


FIGURE 4: MyOcrTool's is illustrated with application's screenshots with different cases such as scanning (a) signboard, (b) book, and (c) restaurant menu.

TABLE 4: A representative image which is shared by many characters having similar meanings.

Character	Pinyin	Meaning	Associative image
何	hé	Why	
吗	má	Question?	
因	yīn	Reason	
由	yóu	Because of	
认	rèn	To recognize	
谁	shéi	Who?	
思	sī	To think	
怎	zěn	How?	
想	xiǎng	To think	
若	ruò	If	
难	nán	Difficult	
如	rú	If	

However, Table 5 lists such characters, where it is not possible for some characters to generate an accurate image to represent them. Because, there are characters without direct or independent image and are difficult to be visualized. Some characters with medium acceptance rate and others with low acceptance rate are shown. This is because, some of the low acceptance rate characters belong to the grammatical structure or modal particles of Chinese writing. Interestingly, these characters are found less in sign boards and restaurants menus. So, failure to generate them accurately

Table 5: Chinese characters and generated images with medium and low acceptance rate.

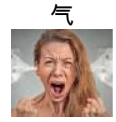
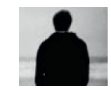
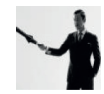
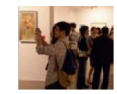
Medium acceptance rate	可	嘻	飞	气	比	希	或	吗	孤	喊
										
Low acceptance rate	了	的	幅	科	苦	斯	展	闻	裹	片
										

Table 6: Evaluation of most suitable distance between the camera and characters.

Scan frame size (cm) (width × height)	Font size	Distance (cm) (vertical direction)
1.5 × 0.5	48	7.1
	26	12.9
	14	19.5

cannot be considered as a major limitation of the system developed. Another solution to this kind of problems is to use the translated captions next to pictures of recognized characters, so that users could perceive the intending meaning of pictures as suggested in [54].

4.3. Testing Based on Different Fonts, Font Sizes, and Varying Distance. While testing the Chinese text which is written in song typeface, bold-face, regular script and imitation song, and their recognition rates were measured. We found that font type has no significant influence on recognition rate. However, if scan box size is fixed, the font size and distance are the two factors that affect each other. In order to measure this, we divided the character size into three levels: large, medium, and small. Characters with large fonts had font size of 48, medium characters had a size of 26, and small characters had a size of 14. After setting the scan box size and font size, we can determine the most appropriate distance required between the camera and the character. The measurement results are shown in Table 6.

4.4. Text from Different Sources. We have tested the accuracy and stability of MyOcrTool in different scenarios, by scanning the text from different sources such as books, warning signs, and restaurant menus. From the test results, we found that MyOcrTool has nearly the same accuracy and stability in these different scenarios. The text recognition rates of software in three different scenarios is shown in Figure 5, where both single word, and long sentences are considered.

Considering the results obtained in the above two cases, the system is showing acceptable performance and can provide better support for Chinese learners to understand the meaning of Chinese characters and text from the perspective of visual information association. The system has a high accuracy rate of about 88%, which can meet the daily learning needs of Chinese learners, but further strengthening of the recognition ability is also necessary.

While testing in two brands of mobile phones (Oppo-R7SM, and Vivo-Y66) it is found that, the average time

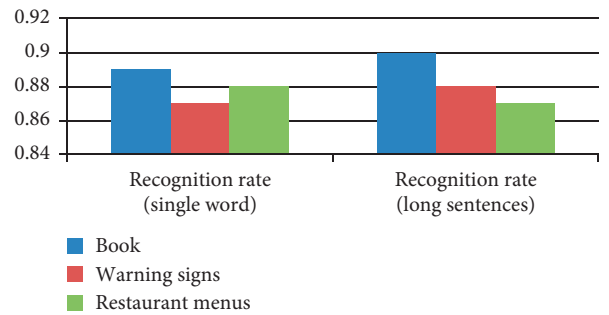


Figure 5: The histogram of recognition rate of MyOcrTool in different text sources.

required to identify a text is 7.8 seconds. The software execution speed is depends on many factors. Firstly, it can be determined according to the specific words. The frequently used words are recognized faster, and if the words are not used frequently the speed will be slower. Secondly, the execution speed is depends on the font type, and number of strokes. It is founds that, if the font design is complicated, the stroke recognition will be slower. For example, comparing two different characters such as "翻" and "天", the latter character is faster than the former. Thirdly, execution speed is also depends on the resolution of camera lens, and higher the resolution, the faster the recognition speed. As the application is developed for handheld devices, there is no need of any Internet access to generate Images, and in the current system all the images are packaged into the program itself. However, having Internet provides future possibilities to integrate with several tools, which we have not considered yet. Moreover, we have also not considered memory space required in the mobile devices, because, as images are of small in size, the entire visualization system takes less system space.

5. Conclusion and Future Work

In this paper, an Android based system named MyOcrTool to capture Chinese characters from different text sources and displaying the visual images associated with them in

real-time along with audio options to listen to the script is developed. MyOcrTool displays the visual image related to a Chinese character in real-time after recognizing the text. With this learners from almost all backgrounds are able to visualize the Chinese text only through scanning in Android based devices. This proves, that, we can conclusively answer the research first research question mentioned in the Introduction section. Moreover, learners do not need to develop the skills such as remembering *pinyin*, or stroke sequences. They can also use this system without even reading or writing Chinese characters, and entering any information to the device to obtain the meaning is absolutely unnecessary. The proposed system is designed for such learners who would like to visualize the daily life Chinese texts for rapid assimilation of meanings behind them. After the experimental evaluation, it is found that, the text recognition rate of MyOcrTool reaches nearly 90%, and the time delay between text recognition, and display of visual image in real-time is less than half seconds. The recognition results obtained conclusively prove that, it is possible to evaluate recognition rate of characters which were extracted from different sources. This answers the second research question.

However, we can also list some of the limitations of this work, and there is scope for further research. Firstly, considering the sources such as text from newspaper articles, as they are written with particular context, and generating image for a sentence is beyond the scope of this work. As the sentences becomes longer, we find that, the characters which can be shown with 100% accurate image are diminishing, because there would be more Chinese pronouns. So, we found testing such features of long sentences, and providing exact recognition considered as part of future work. Secondly, there is scope for improving the text recognition speed through applying better recognition algorithms, and image processing methods. It is also possible to show multiple sequence of images for single character based on context with using pictures using GIF (Graphics Interchange Format) animation to avoid ambiguities in their visualization meanings. In addition, displaying corresponding pictures with translation functions could solve the problems with ambiguous words or pictures. Similarly, in-depth study on handwritten Chinese text recognition, and associative image generation is also necessary. Thirdly, the developed MyOcrTool unable to process the noisy background around a scanned text, which makes the system difficult to identify characters in text sources with messy background. This limitation also reduce the recognition rate and processing speed significantly. Finally, regarding the voice playback feature, more sophisticated and advanced playback engine can be used to make the text-to-speech sound more user friendly, and error-free.

Data Availability

No raw data is required to reproduce the work other than few representative images as shown in Tables 1, 3–5. Three Program Listings are included within the manuscript itself, so that programming support is enclosed to reproduce the application. The representative images are collected by following the earlier work proposed in Reference [7], and Chinese characters are collected from link provided in Reference [2]. The software used to develop the proposed system are obtained from the links provided in References [45, 48].

Conflicts of Interest

The authors declares that there is no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China.

References

[1] X. Yan, Y. Fan, Z. Di, S. Havlin, and J. Wu, "Efficient learning strategy of Chinese characters based on network approach," *PLoS One*, vol. 8, no. 8, Article ID e69745, 2013.

[2] "Learn to read and write simplified Chinese characters," 2018, http://www.learnchineseez.com/read-write/simplified/index.php.

[3] L. H. Wong, C. K. Chin, C. L. Tan, and M. Liu, "Students' personal and social meaning making in a Chinese idiom mobile learning environment," *Educational Technology & Society*, vol. 13, no. 4, pp. 15–26, 2010.

[4] M. You and Y. J. Xu, "A usability testing of Chinese character writing system for foreign learner," *Lecture Notes in Computer Science*, vol. 8013, no. 2, pp. 149–157, 2013.

[5] F. Rosell-Aguilar and K. Qian, "Design and user evaluation of a mobile application to teach Chinese characters," *The JALT CALL Journal*, vol. 11, no. 1, pp. 19–40, 2015.

[6] Omniglot, "Evolution of Chinese characters," 2018, https://www.omniglot.com/chinese/evolution.htm.

[7] L. Rai, T. Yang, Z. Yue, N. Sun, and R. Shadiev, "Visualizing characters as images: understanding Chinese through Internet usage," in *Proceedings Of the 17th IEEE International Conference On Advanced Learning Technologies (ICALT)*, Timisoara, Romania, July 2017.

[8] J. Yang, C. Wang, B. Jiang, H. Song, and Q. Meng, "Visual perception enabled industry intelligence: state of the art, challenges and prospects," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2204–2219, 2021.

[9] K. Sim, J. Yang, W. Lu, and X. Gao, "MaD-DLS: Mean and Deviation of Deep and Local Similarity for image quality assessment," *IEEE Transactions On Multimedia*, pp. 1–12, 2021.

[10] J. Yang, Y. Zhao, J. Liu et al., "No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions," *IEEE Transactions on Cybernetics*, pp. 1–13, 2020.

[11] Z. Lv and D. Chen, "Industrial visual perception technology in Smart City," *Image and Vision Computing*, vol. 105, 2021.

[12] Z. Lv, D. Chen, R. Lou, and H. Song, "Industrial security solution for virtual reality," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6273–6281, 2021.

[13] Z. Lv, "Virtual reality in the context of Internet of things," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9593–9602, 2020.

[14] K. Jung, K. In Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, 2004.

[15] I. Kastelan, S. Kukolj, V. Pekovic, V. Marinkovic, and Z. Marceta, "Extraction of text on TV screen using optical character recognition," in *Proceedings Of the 10th IEEE Jubilee International Symposium On Intelligent Systems And Informatics*, pp. 153–156, Subotica, Serbia, September 2012.

[16] S. Ramiah, T. Y. Liong, and M. Jayabalan, "Detecting text based image with optical character recognition for English translation and speech using Android," in *Proceedings Of the Student IEEE Conference On Research And Development (SCOReD)*, pp. 272–277, Kuala Lumpur, Malaysia, December 2015.

[17] P. Chavre and A. Ghotkar, "Scene text extraction using stroke width transform for tourist translator on Android platform," in *Proceedings Of the International Conference On Automatic Control And Dynamic Optimization Techniques (ICACDOT)*, pp. 301–306, Pune, India, September 2016.

[18] A. Kongtaln, S. Minsakorn, L. Yodchaloemkul, S. Boontarak, and S. Phongsuphap, "Medical document reader on android smartphone," in *Proceedings of the 3rd ICT international senior project conference (ICT-ISPC)*, pp. 65–68, Nakhon Pathom, Thailand, October 2014.

[19] S. Dhakal and M. Rahnemoonfar, "Mobile-basedtext recognition from water quality devices," in *Proceedings of the Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications*, vol. 9411, San Francisco, CA, USA, March 2015.

[20] F. D. Nurzam and E. T. Luthfi, "Implementation of real-time scanner Java language text with mobile vision Android based," in *Proceedings Of the International Conference On Information And Communications Technology (ICOIACT)*, pp. 724–729, Yogyakarta, Indonesia, April 2018.

[21] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 2972–2982, 2014.

[22] J. Tan, X. H. Xie, W. H. Zheng, and J. H. Lai, "Radical extraction using affine sparse matrix factorization for printed Chinese characters recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 3, pp. 211–226, 2012.

[23] X. Wang, H. Du, and X. Wen, "Research on segmentation and recognition of printed Chinese characters," *Journal of Physics: Conference Series*, vol. 1237, 022011 pages, 2019.

[24] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio, "Drawing and recognizing Chinese characters with recurrent neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 849–862, 2018.

[25] Z. Cao, J. Lu, S. Cui, and C. Zhang, "Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding," *Pattern Recognition*, vol. 107, 2020.

[26] H. Chen, B. Hu, X. Yang, M. Yu, and J. Chen, "Chinese character recognition for LPR application," *Optik*, vol. 125, no. 18, pp. 5295–5302, 2014.

[27] M. Zhang, F. Xie, J. Zhao, R. Sun, L. Zhang, and Y. Zhang, "Chinese license plates recognition method based on A robust and efficient feature extraction and BPNN algorithm," *Journal of Physics: Conference Series*, vol. 1004, no. 1, 012022 pages, 2018.

[28] P. Gao, J. Wu, Y. Lin, Y. Xia, and T. Mao, "Fast Chinese calligraphic character recognition with large-scale data," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7221–7238, 2015.

[29] Y. Lu, J. Li, H. Zhang, and S. Lin, "Chinese character recognition of e-commerce platform pictures," in *Proceedings Of the 13th International Computer Conference On Wavelet Active Media Technology And Information Processing*, pp. 28–31, Chongqing, China, March 2017.

[30] Y. Yin, W. Zhang, S. Hong, J. Yang, J. Xiong, and G. Gui, "Deep learning-aided OCR techniques for Chinese uppercase characters in the application of Internet of things," *IEEE Access*, vol. 7, pp. 47043–47049, 2019.

[31] P. Sun, Q. Xie, Z. Wu, X. Feng, J. Cai, and Y. Jiang, "Yi characters recognition based on tesseract-OCR," in *Proceedings Of IEEE 3rd Advanced Information Management, Communicates, Electronic And Automation Control Conference*, pp. 102–106, Chongqing, China, October 2019.

[32] B. Wang, Y. W. Ma, and H. T. Hu, "Hybrid model for Chinese character recognition based on Tesseract-OCR," *International Journal of Internet Protocol Technology*, vol. 13, no. 2, pp. 102–108, 2020.

[33] S. Celaschi, M. S. Castro, and S. P. Da Cunha, "Read it aloud to me," *Universal Access in Human-Computer Interaction. Designing Novel Interactions*, vol. 10278, pp. 260–268, 2017.

[34] G. Chomchalerm, J. Rattanakajornsak, U. Samsrisook, D. Wongsawang, and W. Kusakunniran, "Braille Dict: dictionary application for the blind on android smartphone," in *Proceedings of the 3rd ICT international senior project conference (ICT-ISPC)*, pp. 143–146, Nakhon Pathom, Thailand, October 2014.

[35] G. A. Robby, A. Tandra, I. Susanto, J. Harefa, and A. Chowanda, "Implementation of optical character recognition using Tesseract with the Javanese script target in android application," *Procedia Computer Science*, vol. 157, pp. 499–505, 2019.

[36] G. B. Holanda, J. W. M. Souza, D. A. Lima et al., "Development of OCR system on android platforms to aid reading with a refreshable braille display in real time," *Measurement*, vol. 120, pp. 150–168, 2018.

[37] Y. G. Lee, "Novel video stabilization for real-time optical character recognition applications," *Journal of Visual Communication and Image Representation*, vol. 44, pp. 148–155, 2017.

[38] S.-M. Huang, "Effects of font size and font style of Traditional Chinese characters on readability on smartphones," *International Journal of Industrial Ergonomics*, vol. 69, pp. 66–72, 2019.

[39] A. D. Cheok, Z. Jian, and E. S. Chng, "Efficient mobile phone Chinese optical character recognition systems by use of heuristic fuzzy rules and bigram Markov language models," *Applied Soft Computing*, vol. 8, no. 2, pp. 1005–1017, 2008.

[40] M. C. Lee, S. Y. Chiu, and J. W. Chang, "A deep convolutional neural network based Chinese menu recognition App," *Information Processing Letters*, vol. 128, pp. 14–20, 2017.

[41] M. Koga, R. Mine, T. Kameyama, and T. Takahashi, "Camera-based Kanji OCR for mobile-phones: practical issues," in *Proceedings of the International Conference On Document Analysis And Recognition (ICDAR)*, pp. 635–639, Seoul, South Korea, January 2006.

[42] R. Casey and G. Nagy, "Recognition of printed Chinese characters," *IEEE Transactions on Electronic Computers*, vol. 15, no. 1, pp. 91–101, 1996.

[43] R. Smith, "An overview of the Tesseract OCR engine," in *Proceedings Of the International Conference On Document Analysis And Recognition (ICDAR)*, Parana, Brazil, November 2007.

[44] R. Smith, D. Antonova, and D. S. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in *Proceedings of International Workshop on Multilingual OCR*, Barcelona, Spain, July 2009.

[45] "Using the recyclerview," 2021, https://guides.codepath.com/android/using-the-recyclerview.

[46] H. Jiang, T. Gonnot, W. J. Yi, and J. Saniie, "Computer vision and text recognition for assisting visually impaired people using Android smartphone," in *Proceedings Of the IEEE International Conference On Electro Information Technology*, pp. 350–353, Lincoln, NE, USA, October 2017.

[47] Y. Yi, "The design of 2D bar code recognition software on Android," *Advanced Materials Research*, vol. 442, pp. 453–457, 2012.

[48] Zxing, "Google open source," 2018, https://opensource.google.com/projects/zxing.

[49] G. Chen, "Application of processing techniques from color image to grey image," in *Proceedings Of the 2nd International Conference On Software Technology And Engineering (ICSTE)*, San Juan, PR, USA, October 2010.

[50] C. Kanan and C. W. Cottrell, "Color-to-grayscale: does the method matter in image recognition?," *PLOS One*, vol. 7, no. 1, Article ID e29740, 2012.

[51] S. S. Lokhande and N. A. Dawande, "A survey on document image binarization techniques," in *Proceedings Of the 1st International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pp. 742–746, Pune, India, July 2015.

[52] S. H. Shaikh, A. Maiti, and N. Chaki, "Image binarization using iterative partitioning: a global thresholding approach," in *Proceedings Of the International Conference On Recent Trends In Information Systems(ReTIS)*, pp. 281–286, Kolkata, India, December 2011.

[53] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Efficient and effective OCR engine training," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, no. 1, pp. 73–88, 2020.

[54] R. Shadiev, T.-T. Wu, and Y.-M. Huang, "Using image-to-text recognition technology to facilitate vocabulary acquisition in authentic contexts," *ReCALL*, vol. 32, no. 2, pp. 195–212, 2020.

Hindawi
Complexity
Volume 2021, Article ID 6660572, 22 pages
https://doi.org/10.1155/2021/6660572

Research Article

A Methodology to Determine the Subset of Heuristics for Hyperheuristics through Metalearning for Solving Graph Coloring and Capacitated Vehicle Routing Problems

Lucero Ortiz-Aguilar,[1] **Martín Carpio**,[1] **Alfonso Rojas-Domínguez**,[1] **Manuel Ornelas-Rodriguez**,[1] **H. J. Puga-Soberanes**,[1] **and Jorge A. Soria-Alcaraz**[2]

[1]*Tecnológico Nacional de México, Instituto Tecnológico de León, Guanajuato, Mexico*
[2]*Department of Organizational Studies, University of Guanajuato, Guanajuato, Mexico*

Correspondence should be addressed to H. J. Puga-Soberanes; pugahector@yahoo.com

Received 6 October 2020; Revised 15 March 2021; Accepted 12 April 2021; Published 26 April 2021

Academic Editor: Abd E.I.-Baset Hassanien

In this work, we focus on the problem of selecting low-level heuristics in a hyperheuristic approach with offline learning, for the solution of instances of different problem domains. The objective is to improve the performance of the offline hyperheuristic approach, identifying equivalence classes in a set of instances of different problems and selecting the best performing heuristics in each of them. A methodology is proposed as the first step of a set of instances of all problems, and the generic characteristics of each instance and the performance of the heuristics in each one of them are considered to define the vectors of characteristics and make a grouping of classes. Metalearning with statistical tests is used to select the heuristics for each class. Finally, we used the Naive Bayes to test the set instances with k-fold cross-validation, and we compared all results statistically with the best-known values. In this research, the methodology was tested by applying it to the problems of capacitated vehicle routing (CVRP) and graph coloring (GCP). The experimental results show that the proposed methodology can improve the performance of the offline hyperheuristic approach, correctly identifying the classes of instances and applying the appropriate heuristics in each case. This is based on the statistical comparison of the results obtained with those of the state of the art of each instance.

1. Introduction

In Computer Science, a heuristic is a technique designed to solve a problem when classical methods fail to find an exact solution or when they are too slow. Currently, there is great interest from the scientific community in offering ad hoc heuristic solutions for real-world optimization problems. To achieve this, it is necessary to have a priori knowledge of the problem and, often, computationally efficient solutions are produced in a reasonable time. However, the no free lunch theorem mentions [1] that no methodology or algorithm can solve all problems; that is, ad hoc heuristics are usually not generalizable, and they do not always work well when applied to other problems even when they share some similar characteristics. This fact has led research efforts towards the development of general-purpose search methodologies known as hyperheuristics, whose main characteristic is that they are independent of the problem domain.

Hyperheuristics can be classified according to their learning methods, such as no learning, online learning [2], and offline learning [3]. In the context of combinatorial optimization, hyperheuristics are defined as "heuristics to choose heuristics" [4], or as "an automated methodology for selecting or generation heuristics to solve computational search problems," [4]. According to Pillay in [2], the generality of a hyperheuristic can be seen from three levels: a generalization on instances of problems, generalization for a particular problem, and a generalization focused on different types of problems, the latter being high level. Some variations of hyperheuristics depend on the type of learning used (e.g., with online learning [2] and offline learning [3]) or the nature of the heuristics.

One of the main problems in hyperheuristics is to propose methodologies that allow generating and/or selecting the minimum set of heuristics that perform well for the problem at hand and this heuristic set is usually selected by expert researchers in the field [2]. In order to automatically select the best heuristic that performs best for the problem, an approach called metalearning was proposed [5, 6] and its use in hyperheuristics can be found in Amaya et al. [7]. Likewise, the different meanings used and taxonomy for each interaction between metalearning and optimization were studied by Song et al. [8].

As there is no methodology or algorithm that can solve all problems, our objective is to base ourselves on information about the problem and the performance of the algorithms to provide this knowledge to hyperheuristics. Metalearning generates metaknowledge, and we use this to select better heuristics for solving problems. With this approach, we pretend to propose methodologies that are like the intelligence of humans. Humans could learn from problems, their characteristics, their variables, and their restrictions, and after elaborating an analysis or discernment, the "human expert" proposes the best tool and solves the problem.

In this paper, we propose a methodology to determine a subset of heuristics for hyperheuristics through metalearning and partition for solving different problems (described below) without ad hoc adjustments by providing information about the problem and the performance of the heuristics to the hyperheuristic. It is well known that the correct characterization is the key to selecting the best heuristic [6]. Consequently, this affects the hyperheuristic design, and that is why in our approach we decided to use offline learning. Metalearning consists of two basic parts, the metafeatures and the metalearner; the first are generated from the information of the problem and the solution algorithms, while the second uses a grouping technique. Our methodology extends beyond the classic metalearner approach, and we apply nonparametric statistical tests to determine which heuristics will provide the same performance if the full set of heuristics is applied.

In order to test our proposal, we used two different well-known problem domains: capacitated vehicle routing problem and graph coloring problem. The capacitated vehicle routing problem (CVRP) has different restrictions such as minimizing distance, time, capacity, and delivery. This problem aims to find the subtour of n cities, without repeating two cities on the same tour or different tours. In the CVRP state of the art, there exist different variants that consider this basic definition and extra restrictions. On the other hand, the graph coloring problem consists of labeling each vertex of a given graph with k-colors and it is a well-known problem, which has been solved by exact methods, heuristics, metaheuristics, and hyperheuristics. Although each problem can be solved by ad hoc heuristics, to date, there is no general methodology capable of solving all variants of both problems. The use of partitions for constraint satisfaction problems, such as university timetabling and VRP, has had good results [9, 10]. Although there are taxonomies of the said problems, characterizing and classifying the instances under a hyperheuristic context with metalearning with a statistical test is an approach that has not been explored in the literature.

Finally, it is worth mentioning that to do a better design or choice of hyperheuristics and predict which is the best algorithm according to the classification of the previous instance, we propose to use metalearning with statistical analysis of the heuristics, which will allow improving these points. Our proposal also provides information that allows us to understand the performance of heuristics and hyperheuristics in the problem of interest.

The remaining content of this article contains a description of related work in Section 2, which covers a review on heuristics, hyperheuristics, and metalearning. Problem definitions and theories related to heuristics are reported in Section 3. Sections 4 and 5 present the proposed methodology. The found results and findings including performance comparison are described in Section 6. Finally, concluding remarks are presented in Section 7.

2. Related Work

In this section, we will give a view of the heuristic and hyperheuristic algorithms. Moreover, we will define some basic concepts of metalearning. Finally, we will give a discussion of the pros and cons of the presented methods.

2.1. Heuristics. We made an extensive review of different heuristics applied to CVRP and GCP. We selected a total of 11 heuristics that, after a previous experimental analysis, those heuristics apply to both problems and we list them below.

K-flip or *K-opt* heuristic was proposed by Lin and Kernighan in [11] for the travel salesman problem (TSP). This heuristic was based on the general interchange transformation, i.e., a city must change its position with another city on the same tour. Besides, this heuristic is one of the most popular for TSP [12]; it has been applied in other problems such as planar graphs, unconstrained binary quadratic programming, and the study of its complexity in SAT and MAX-SAT. The two-point perturbation is a case of *k-flip*, and we give a detailed description and algorithm of these heuristics in the following sections.

The *k-swap* heuristic is similar and frequently confused with *k-flip*. The *K-swap* heuristic improves its performance as a perturbation move when it uses two or three movements [13].

The move to less conflict heuristic, also known as minimizing conflicts, was proposed by Minton et al. [14]. The minimizing conflict heuristic has been applied to different areas in Computer Science such as hyperheuristics, graph coloring problems, pickup-and-delivery problems, and scheduling problems. The move to less conflict heuristic is a variant of the first fit, and the only difference is that the first one takes a random variable and changes its value for another that generates the least cost.

The *first-fit* heuristic was proved by Baker [15] for the bin-packing problem. On the other hand, in recent decades,

this heuristic was applied to best-known problems such as bin packing, virtual machine relocation problem, and cutting stock. A remarkable variety of heuristics is *worst fit* which was studied by Baker [15] and Csirik [16], in particular, its application to bin-packing problem.

Soria-Alcaraz et al. [17] proposed three heuristics for university course timetabling best single perturbation (BSP), static-dynamic perturbation (SDP), and double dynamic perturbation as part of the pool low-level heuristics for hyperheuristic. Moreover, these heuristics were applied to the VRP in later research [18].

2.2. Hyperheuristics. We focused on offline learning hyperheuristics selection with perturbation heuristics, whose aim is to gather knowledge in the form of rules or programs, from training set instances. Usually, the offline selection hyperheuristics belong to machine learning methods, which are trained to create a tuned methodology for a problem domain [3]. Yates and Keedwell [19] demonstrated that subsequences of heuristics were found in the offline learning database that is effective for some problem domains. They used the Elman network to compute sequences of heuristics which were evaluated on unseen *HyFlex* example problems, and the results obtained are capable of intradomain learning and generalization with 99% confidence.

One of the crucial issues in hyperheuristics design is the quality and size of the heuristic pool [20]. Soria-Alcaraz et al. [20] proposed a methodology using nonparametric statistics and fitness landscape measurements for hyperheuristics design. This methodology was tested on course timetabling and vehicle routing problems; their hyperheuristic proposal had a compact heuristic pool and competed with some traditional methods in course timetabling. In the course timetabling problem, they obtained five best-known solutions of 24 PATAT instances [21]. Finally, a recent report by Amaya et al. [7] documented a model for creating selection hyperheuristics with constructive heuristics. The effectiveness of the model proposed by Amaya depends on the delta's values used, which is useful with higher deltas.

2.3. Metalearning. The importance of metalearning, machine learning, and optimization has been studied by Song et al. [8]. The metalearning aim may concern accumulating and adapting experiences on the performance of multiple applications of a learning system. The metalearning field is also known as "*learning to learn*" [22] and it brings systems that can help by searching patterns across different tasks to control the process of exploiting cumulative expertise. The metalearning concept has been present in the field of heuristics and metaheuristics for TSP [23], the quadratic assignment problem, and hyperheuristics.

On the other hand, Gutierrez-Rodríguez et al. [23] used VRP with time windows and proposed a methodology based on metalearning to select the best metaheuristic for each instance. Besides, their proposal shared and exploited an offline scheme for the instant solutions of academies and industry. Their main contributions were to propose a set of features for characterizing VRPTW instances and design a classification process that predicts the most suitable metaheuristic for each instance. Nevertheless, they assumed that the solutions of the set instances could be stored, shared, and exploited in an offline scheme for predicting good solvers for new unseen instances.

The aim of this paper is not to present a survey on heuristics or hyperheuristics; our proposal is slightly different. Our proposal considers some vital aspects of the research, including the ones from Yates and Keedwell [19]. We took the offline hyperheuristic approach from Soria-Alcaraz et al. [20] and the statistical approach to selecting a pool heuristic from Kanda et al. [5]. The offline hyperheuristic approach is an effective and popular method in the machine learning area [8]. On the other hand, the statistical approach to selecting a pool heuristic is a useful and reliable method because it takes statistical information from the input data.

3. Combinatorial Problems

Our methodology is a general approach to competitive performance across several classes of problems. Thus, we used two problem domains: graph coloring and vehicle routing problems. In the following sections, we will review the formal definition of each of these problems as well as their benchmark instances.

3.1. Graph Coloring Problem. The graph coloring problem demonstration as an NP-hard problem was proposed by Karp [24]. According to [25], a formal vertex-coloring problem of a graph $G = (V, E)$ is a function c: $V \longrightarrow \mathbb{N}$, in which any two incident vertices $u, v \in V$ are assigned different colors, that is, $\{u, v\} \in E \Rightarrow c(u) \neq c(v)$, and E is a finite set of unordered pairs of vertices named *edges*, where the function c is the coloring function and a graph G for which there exists a vertex-coloring which requires k colors is called *k-colorable*. The coloring function induces a partition of the graph G into independent subsets $V_1, V_2, \ldots, V_k$, where $V_i \cap V_j = \varnothing$ and $V_1 \bigcup V_2 \bigcup \ldots \bigcup V_k = V$. The benchmark instances can be found in http://mat.gsia.cmu.edu/COLOR/instances.htmlmat.gsia.cmu.edu. The above lets the partitioning methodology work on the input design and it is possible to avoid the ad hoc modifications to the heuristics since it will only pass a different objective function that adequately evaluates the instances of this problem.

3.2. Capacitated Vehicle Routing Problem. The capacitated vehicle routing problem (CVRP) is a variant of VRP [26]. In this problem, we have an undirected graph G, m vehicles, Q capacity, and a set of cities $C = \{c_0, c_1, \ldots, c_n\}$. Formally, the city c_0 is the *depot* and each vehicle must visit these cities starting from the *depot* and coming back to this. Alba and Dorronsoro [27] define a distance or travel time matrix $D = (d_{i,j})$ between cities c_i and c_j. Each city c_i has a demand of things q_i. We denote it as a route $R = \{\overrightarrow{r}_0, \overrightarrow{r}_1, \ldots, \overrightarrow{r}_m\}$, and $\overrightarrow{r}_i$ is a permutation of the cities, starting and finishing at the depot c_0. For each route, $\overrightarrow{r}_i \cap \overrightarrow{r}_j = c_0$ with $i \neq j$. The cost of a problem solution is the sum of the costs of each route of R as

$$F(R) = \sum_{i=1}^{m} Cost\left(\overrightarrow{r}_i\right), \quad (1)$$

$$Cost\left(\overrightarrow{r}_i\right) = \sum_{j=0}^{k} d_{j,j+1}, \quad (2)$$

where k is the total of vehicles. This problem aims to determine for each vehicle the lowest cost (see equations (1) and (2)) tour or distance or travel time, considering the max capacity. Nota bene the hard constraints are the capacity of each vehicle and two vehicles cannot visit the same city. The CVRP has several constraints and a specific formal definition of this problem. These two characteristics let us apply our methodology with a design cities partition, where each vehicle is related to one part. As heuristics work with solutions that are already complete and respect important restrictions such as capacity, if any movement violates or exceeds this capacity, that solution is penalized in the objective function.

4. Methodology

For our methodology, it is important to know since the constraint modeling phase of the problem can be solved by partitions. The API-Carpio methodology and the methodology proposed by Soria-Alcaraz et al. [28] let us transform the instances of the problem with their restrictions, into inputs to apply the proposed methodology. The MMA matrix that is generated by applying the API-Carpio methodology lets us visualize the hard restrictions of the problem and evaluate the costs of visiting cities or nodes. For the soft constraints of the problem, the methodology proposed by Soria-Alcaraz et al. [28] is to be considered in the list of restrictions.

In this section, we describe the methodology to model the input data problem information used for the experimentation for two combinatorial problems based on the API-Carpio methodology. We integrated the API-Carpio methodology [29] and the methodology of the design proposed by Soria-Alcaraz et al. [28].

4.1. API-Carpio Methodology. This methodology is used to solve the university course timetabling problem and it considers three factors: students, teachers, and institutions (infrastructure). The methodology uses several structures for the equations previously described. One of the most important structures of this work is MMA. This matrix is constructed with information on the cities or nodes. For graph coloring, we use the information of the adjacency matrix, while for CVRP it is considered the cost matrix. Table 1 shows an example of an MMA matrix. The algorithm to construct this matrix is given in [29].

4.2. Methodology of Design. This methodology was extended from the proposal by Carpio [29] and their formal definition was proposed by Soria-Alcaraz et al. [28]. The methodology of design by Soria allows us to consider the objectives of *course timetabling* and to satisfy the different restrictions, by converting these to lists of time and space restrictions it is seeking to minimize student conflict.

Table 1: MMA matrix.

City/node	N_1	N_2	...	N_{n-1}	N_n
N_1	**10**	99	...	97	137
N_2	99	**101**	...	80	96
...	...	...	...	...	...
N_{n-1}	97	80	...	**35**	10
N_n	137	96	...	10	**68**

To use this methodology, two structures are used to consider restrictions and variables: MMA matrix and LPH. The LPH has information about the possible restrictions that can be assigned to each node or city. An example of this list can be found in Table 2. The list shows in each row the number part, i.e., the node N_2 can be assigned in parts 1, 2, 4, 5, 7, or 8, but not in 6. The algorithm for generating artificial instances of LPH can be found in Ortiz-Aguilar [30].

5. Metalearning for Selecting a Subset of Heuristics for Hyperheuristics

According to Brazdil et al. [22], the approach of metalearning (ML) is to help the selection of an algorithm for a set of instances with metadata. According to Alpaydin [31], the metalearning aim is to find the best classifier for a set of data and to find the best classifier for the characterization when the data are considered. In our proposal, the data are associated with the problem instances of different problems, and the classifier is associated with the set of heuristics. Our objective is to be able to select the best set of heuristics for a hyperheuristic in a set of instances. In Figure 1, we show a diagram of the metalearning processes to obtain metaknowledge for the selection of heuristics (diagram modified from Brazdil and Giraud-Carrier [32] to our methodology).

In this work, we use metalearning to select a set of heuristics for hyperheuristics in a dataset. We named the set of characterized instances of the two problem domains as the "metacharacteristics" and the model that maps each instance to the corresponding group of heuristics for hyperheuristics the "metalearner." In this case, the metalearner selected is the *K-means* algorithm. The methodology proposed in this article consists of 5 steps in the metalearning stage:

(i) Step 1: obtain the set of instances to be worked on. In this case, we have as a criterion to select those instances that are susceptible to being resolved by partitions.

(ii) Step 2: evaluation and extraction of characteristics of the instances. In this step, the characteristics of the heuristics and the instances are generated. Heuristics that apply to both problems are selected, this task becomes simple with the use of the partitioning methodology, and this is because it allows working always with generic inputs where the variables and restrictions are modeled. Later, heuristics work with these generic inputs and solutions that only have a fitness function corresponding to their problem (where the objectives are evaluated).

Table 2: LPH list.

Node or city	Restrictions							
N_1	1	2	3	4	5	6	7	8
N_2	1	2	3	4	5		7	8
...	...	...	...	...	...	...	...	...
N_{n-1}	1	2	3	4	5	6	7	8
N_n	1	2	3	4	5	6	7	8

$N_1 ... N_n$ represents the node or cities.

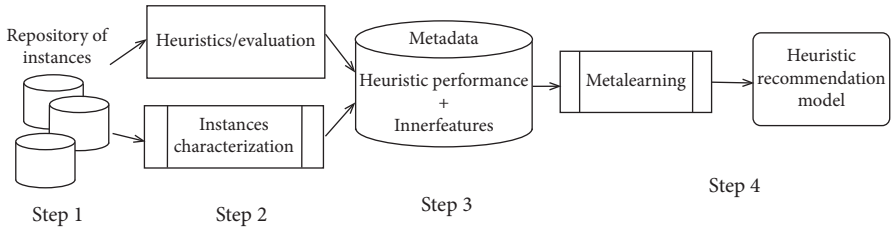


Figure 1: Metalearning to obtain metaknowledge for heuristic selection.

(iii) Step 3: generation of metacharacteristics. Based on the characteristics of each problem and the performance of the heuristics applied to all instances, we generate vectors of characteristics that will be our metadata.

(iv) Step 4: metalearning and the recommended model of heuristics. In the state of the art, research is limited to applying only a clustering technique for the recommendation of the algorithm model. We propose to incorporate a statistical analysis together with the clustering algorithm to improve the design of the basic subset of heuristics.

5.1. Problem Definition of Metalearning for Heuristic Selection. Consider a problem P_i that belongs to the problem set (GCP). Let $H_i = \{h_{i,1}, h_{i,2}, \cdots, h_{i,n_i}\}$ be a subset of low-level heuristics, which are used in the state of the art, to solve the problem P_i. We denote as RS_{H_i} a random selection of heuristics H_i, to be applied in the solution of P_i, where $|RS_{H_i}| = |H_i|$, with RSH_i to a reduced set of heuristics, that is, $RSH_i \subseteq H_i$.

Let $V_0 = \{\lambda\}$, where λ represents an empty string. We defined recursively $V_{n+1} = \{wh_{i,m} | w \in V_n; h_{i,m} \in H_i\}$ where $n \geq 0$ and $m \in \{1, 2, 3, \cdots, n_i\}$. Then V_n represents the set of all strings of length $n \in \mathbb{Z}^+ = \{0, 1, 2, 3, \cdots\}$, formed from the symbols in H_i. So Kleene's closure from H_i is [33]

$$H_i^* = \bigcup_{n \in \mathbb{Z}^+} V_n = V_0 \bigcup V_1 \bigcup V_2 \bigcup V_3 \bigcup \cdots. \quad (3)$$

When V_0 is omitted at the junction, we get the Kleene Plus closure H_i^+:

$$H_i^+ = \bigcup_{n \in \mathbb{N}} V_n = V_1 \bigcup V_2 \bigcup V_3 \bigcup \cdots. \quad (4)$$

In other words, H_i^+ is the collection of all possible nonempty strings of finite length generated from the symbols in H_i.

Let HH be a heuristic selection hyperheuristic with offline training, where the training considers the set H_i. After training, the HH provides a methodology $M_{H_i} \in H_i^+$, with the best order of application of low-level heuristics, which we will denote by $BM_{H_i} \in H_i^+$, in the solution of the problem P_i, which improves the performance of the application of a RS_{H_i}, where $|BOS_{H_i}| \geq |H_i|$ [17].

We take two problems P_i and P_j that belong to the problem set (GCP) that comply with level 3 of generality proposed by Pillay [2] and are susceptible to being solved by partitions. Let $H_i = \{h_{i,1}, h_{i,2}, \cdots, h_{i,n_i}\}$ and $H_j = \{h_{j,1}, h_{j,2}, \cdots, h_{j,n_j}\}$ be subsets of low-level heuristics, which are used in the state of art, to solve the problems P_i and P_j, and $UH = \{H_i \bigcup H_j\}$. We denote as RS_{UH} a random selection of the heuristics UH, to be applied in the solution of P_i and P_j, where $|RS_{UH}| = |UH|$, with $RSUH$ to a reduced set of heuristics, that is, $RSUH \subseteq UH$.

Heuristic selection hyperheuristic is denoted as HH, with offline training, where the training considers the UH set. After training, the HH provides a methodology $M_{UH} \in UH^+$, with the best order of application of low-level heuristics, which we will denote by $BM_{UH} \in UH^+$ in solving the problems P_i and P_j which improve the performance of the application of a simple RS_{UH}, where $|BM_{UH}| \geq |UH^+|$.

Our objective is to propose a methodology that provides the HH, with a reduced subset $RSUH \subseteq UH$ for its training, such that the HH provides a methodology $M_{RSUH} \in RSUH^+$, with the best order of application of the reduced set of heuristics $RSUH$, which we will denote by $BM_{RSUH} \in RSUH^+$, in solving the problems P_i and P_j respectively, which equal the performance than with the application of the methodology $M_{UH} \in UH^+$ where $|BM_{RSUH}| \geq |RSUH|$.

To solve the problem, the independent application of each of the heuristics of H_i and H_j was proposed, measuring their performance in solving the problems P_i and P_j. Apply statistical tests to contrast the performance of the independent heuristics and thereby discriminate from each set H_i and H_j those heuristics that obtained the lowest performance.

Next, we will focus on describing the stage of extracting characteristics from heuristics and instances. Our methodology improves the metalearning stage (step 4) with the application of nonparametric statistical tests to determine which heuristics are the ones that will provide the same performance if the full set of heuristics will be applied. This means that, if there are heuristics that are redundant, it is possible to leave them out and consider only those that enhance the speed of the search for solutions to the problem. The metalearning process proposed in this work to select the pool of heuristics includes the following steps:

(1) The problems will be the source of information for the basic features.

(2) Given a set of instances denoted as I, for each of the instances I_i apply a number k of times the heuristic H_j. The results will be the inner features. For the CVRP, a greedy heuristic is used, which will allow us to build feasible solutions to the problem. For graph coloring, we will initialize with a random construction heuristic. The next step is to apply the heuristics. It is possible after this step that the instances can be solved to the best solution due to their complexity. This means that it is possible to avoid executing a complete and expensive computation process when solving problems with the application of a simple heuristic.

(3) With the information obtained from points 1 and 2, feature vectors will be formed that will be our metadata.

(4) For a better treatment of the metadata from the previous step, the following steps are carried out:

(a) The patterns that will be used in the k-means will be scaled with the following formula [34]:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{5}$$

where $x \in \mathbb{R}$ are the values of the original variables (features).

(b) Generate the pattern based on the *inner* and *basic features* per instance. The *basic features* will be the problem information and the *inner features* will be the fitness obtained by the heuristics applied k times to the problem.

(5) Pass the feature vectors to the clustering algorithm to form classes. According to Brazdil and Giraud-Carrier [32], the k-means is a simple learning method, which we apply to carry out the grouping of instances in classes. To determine which are the number of classes to form, we use Sturges rule, since, with data that are the potency of 2, it is approximated in a good way. Determine the subclasses with Sturges rule [35] with

$$c = 1 + \frac{log(T)}{log(2)}, \tag{6}$$

where T is the total amount of data.

As the distance metric for the K-means, we use the Mahalanobis distance; this distance has properties such as being invariant to scale by nonsingular linear transformations. An in-depth study of different metrics [36, 37] will be a specific job to investigate whether it can improve the performance of the proposed methodology.

(6) Label each pattern according to the group number in which each pattern (instance) was classified.

(7) Apply again the three statistical tests to the results of heuristics per problem, according to the formulas in [38]. The test ranks 1 to the best performing heuristic, 2 to the second-best, and n to the worst-performing heuristic. From these tests, we will take the range of the heuristics and the range will now be considered as inner features. With this information and the class label, they will now form patterns.

(8) Determine a cutoff point for each class based on the range, and in this case, it will be the average of the minimum and maximum range. Choose those heuristics that pass the cutoff criterion to be part of the minimum set.

(9) The output will be the minimum set of heuristics per class.

This process is shown in a specific way in Figure 2. The two important aspects of metalearning in our work are heuristics and metacharacteristics, which are low-level heuristics and metafeatures.

5.2. Low-Level Heuristics. An important part of the hyperheuristic approach is the selection of the heuristic set. This article proposes to extract information from heuristics and problems to generate the metafeatures [32]. This lets us improve the design and testing of the hyperheuristic algorithm. The goal in this stage is to generate metafeatures in which the heuristics can have a better performance individually for all problem instances. This improves the next part in which the hyperheuristic must choose the sequence application for each heuristic and it uses a minimal pool of heuristics which is a fundamental part of it [2]. For all instances, it will be applied k times for each heuristic. The heuristics were applied to the two problems and their respective instances were as follows:

(1) *K-Flip/Simple Random Perturbation (SRP)* (H_1). The heuristic changes the value of one or more variables (in some cases k) to another feasible value. The GCP aim is to change the color of a certain node to another [39]. Finally, for CVRP, the movement implies changing a city to another specific vehicle [12].

(2) *K-Swap/Kempe Chain Neighborhood/S-Chain Neighborhood* (H_2). It must be selected two or more varieties and then interchange their values among them when possible; otherwise, the change is not made. We exchange the color between nodes previously selected by GCP. This heuristic is using in

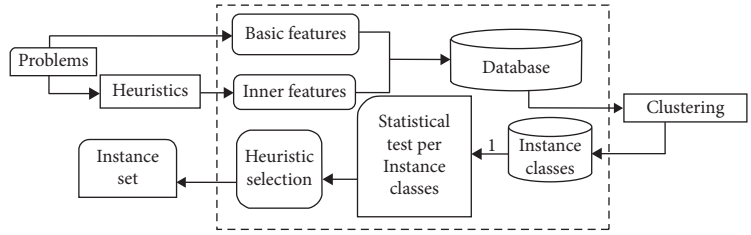


Figure 2: Metalearning process for the heuristics set selection.

works related to TSP or CVRP, also called k-interchange [40, 41].

(3) *Best Single Perturbation (BSP)* (H_3). This heuristic chooses a variable according to the list of hard restrictions (LPH) and changes its value. This exchange produces a better cost or in the worst case, the same cost [17]. Next time this heuristic is going to apply, the next variable will be chosen according to the next position of the last variable which was modified. The next node which must change color will be selected according to the last variable chosen for the graph coloring problem. The CVRP must be changing the city of the vehicle to another vehicle.

(4) *Static-Dynamic Perturbation (SDP)* (H_4). It is also known as statically dynamic perturbation (SDP). It is based on the variable selection with a probability distribution of the frequency in the last k iterations. This heuristic chooses a variable and changes its value randomly [17]. The variables with fewer changes will have a higher probability to be selected. Applied to GCP, it would be a node with fewer color changes, and for the CVRP, the city has moved a few times to another vehicle.

(5) *Two Points Perturbation (2pp)* (H_5). It is also known as *k-opt*, and it is a particular case of the K-swap with a value of $k = 2$.

(6) *Double Dynamic Perturbation (DDP)* (H_6). This heuristic is based on the SDP, this receives a solution, and it modifies the value of a variable concerning a probability distribution. The difference is that a copy of the initial solution is kept and, in the end, the best of the two solutions is returned [17].

(7) *Move to Less Conflict (MLC)* (H_7). This selects a random variable, and it assigns to a part of the value which generates the least cost [18]. In GCP, the color changes according to another which improves the fitness, and in CVRP, the city is moved to another vehicle where the total distance of the route is minimized.

(8) *Min-Conflicts* (H_8). The heuristic selects a random variable, and it assigns to a part which generates the cheapest cost [18]. In GCP, the heuristic must change color from the selected node to another, which improves the result. For CVRP, the selected city with a lower cost in which it must minimize the total distance of the route.

(9) *First-Fit* (H_9). It changes the value of a variable to another, which is the least repeated in other variables [18], i.e., in CVRP, the heuristic will take a city and it will change it to the vehicle that has fewer cities in its route. For the GCP, it will select a node and it will assign the color that is least repeated.

(10) *Worse Fit* (H_{10}). It assigns the most repeated value if possible, without violating the hard constraints on a randomly selected variable [42]. For GCP and CVRP, we assign a node or city to the most repeated timeslot, color, or vehicle.

(11) *Burke–Abdullah (BA)* (H_{11}). This heuristic was proposed by Abdullah et al. [43], in which it chooses a variable applying Fail-First or Brelaz Heuristic [44] and its value changes according to the one that has obtained better performance by applying the following algorithms: minimum conflict, random selection, sequential selection, and least constrained.

5.3. Metafeatures. The description and generation of characteristics permit differentiation into at least two groups of instances within the same problem class. We used the terms of basic feature and *inter feature* based on the proposal conducted by Gutierrez-Rodríguez et al. [23]. As basic features, these are given by the problem, e.g., the number of nodes, colors, vehicles, and so on, depending on each problem information. For both classes of problems, the number of different *basic features* is summarized in Table 3.

The fitness performance values of all heuristics are the inner feature key. Finally, the pattern per instance is *basic feature + inner feature*. The final pattern is shown in Table 4. For example, instance 1 has a pattern (3, 3, 8, 50, 3, 2, 1, 4), and the number of *inter feature* is according to the pool heuristics (eight features for the given example).

6. Methodology for Determining a Subset of Heuristics

In this section, we propose a new approach for selecting and determining a subset of heuristics to solve GCP and CVRP instances. We describe our methodology in the next steps and the graphical representation of our methodology is shown in Figure 3.

(1) *Variables and Problem Restrictions Identification.* First, the variables and restrictions of the problem are identified according to the problem aims or

Table 3: Basic features for CVRP and graph coloring.

Problem	Min partition	Max partition	Variables	Edges
VRP	Vehicles	Vehicles	Cities	Cities connections
Graph coloring	Lower bound	Upper bound	Nodes	Edges

Table 4: Characteristics for instances.

Instance	Lower bound	Upper bound	Nodes	Edges	H_0	H_1	H_2	H_3
1	3	3	8	50	3	2	1	4
2	5	10	6	36	4	3	1.5	1.5
3	3	8	25	80	3	1	2	4
4	10	10	50	80	2	4	3	1

objectives. To model the GCP, the values in the MMA matrix represent the weights of the edges of nodes. If there is a zero in a certain position (x, y) in the matrix, this represents no connection between those nodes. For graph coloring, each node is colored considering that the adjacent nodes do not have the same color. CVRP is aligned with our methodology due to its aim seeking to get subroutes in which the tour cost (subgroup) must be the minimum or the cheapest.

(2) *Problem's Restriction Modeling*. In both problems, we must design a partition of nodes or cities. First, it is necessary to model the restrictions for each variable in an LPH, e.g., a node cannot be colored by a specific color or a restricted city for a tour. Then, it must design the MMA which represents the edge or connection weight between nodes or cities. In GCP, the adjacency matrix corresponds to MMA, and in CVRP, the MMA matrix will be the matrix that has the distances of the node to node. For GC, our LPH is constructed based on the number of colors in which the nodes can be labeled. In case the problem restricts colors to five, the list will be like the one shown in Table 2. Similarly, this list will be built for the CVRP, where the number of vehicles is the number of *parts* that should be represented on the list (see Table 2). For the problems used in this work, it was not necessary to elaborate additional structures for soft restrictions. Besides, for an extensive review and how to model additional restrictions, the research proposed by Ortiz [10] details all possible cases and different features.

(3) Apply the metalearning process described in Section 5.1

(4) Separate the patterns (step 6) into training and test sets to proceed to the classification phase. It is important to consider at least one pattern of each class in the test set.

(5) Use the classifier on the training set to make necessary adjustments to it. After describing and getting all pattern characteristics per instance, the next step is training and testing all instances by a classifier. For our approach, we prefer to use a simple classifier as Bayesian because our objective was not to compare the performance between classification algorithms or to design ad hoc classifiers for our research. The NBC simplifies learning by assuming per class that all features are independent [45]. In our methodology, we assume that each heuristic performance is independent because we applied each heuristic to independent experiments. In the previous stage, each experiment must be run with only one heuristic, and thus, we did not apply two or more heuristics at a time. Finally, all features in the created dataset instances were normalized before applying the classifier.

(6) Finally, the set of test instances will use the classifier to assign a "class" and solve it with its corresponding set of heuristics.

6.1. Designing and Testing the Hyperheuristic Offline Learning with K-Folds. To choose the minimal set of heuristics and design the hyperheuristic for each class in more detail, our methodology considered the hyperheuristics with offline training as it has demonstrated good results for constraint satisfaction problems in terms of generality solution [3]. A random constructive heuristic was used to generate solutions to our problem of GCP, and for CVRP, a greedy algorithm was used. A selection hyperheuristic algorithm has three components: the pool of operators (low-level heuristics), a high-level search strategy, and a control mechanism to select the operator, which will be applied at each search step.

6.1.1. High-Level Search Strategy. The iterated local search algorithm was used as a high-level search strategy. This metaheuristic was proposed by Lourenço et al. [46] and it is constructing a sequence of solutions generated by an embedded heuristic. The generated solutions could be better if they were only constructed randomly. The essence of this algorithm is to intensify an initial solution, exploring neighboring solutions to it. The algorithm is shown in Algorithm 1, which was taken from El-Ghazali [47]. In the field of hyperheuristics with offline learning, it refers to the fact that the high-level search strategy searches for a methodology (a sequence of heuristics) that solves a set of instances and then applies it to a given set of instances, in contrast to online learning, which refers to the construction of a given sequence of heuristics as the instances are presented.

6.1.2. Selection Operator. In the perturbation phase (step 4 in Algorithm 2), it is necessary to choose a variable following a probability distribution based on the frequency of variable selection in the last k iterations. This simple heuristic allows

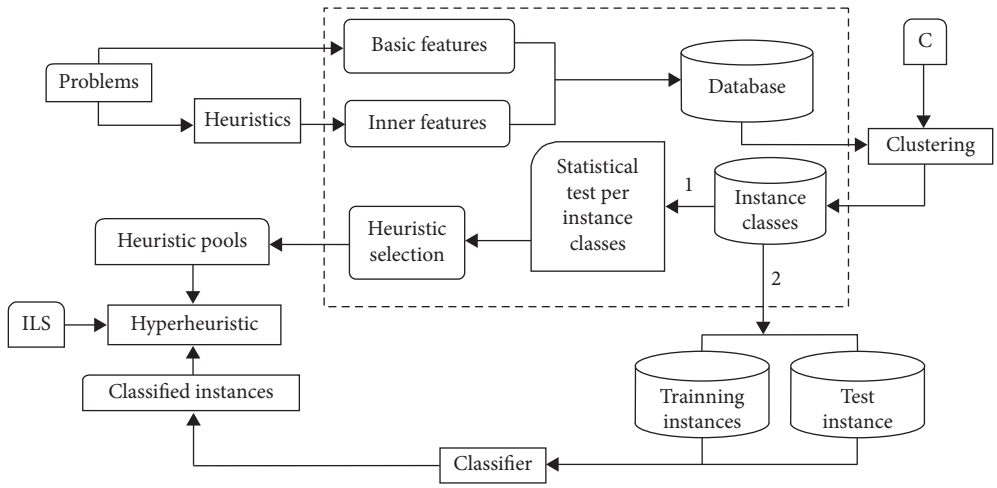


Figure 3: Methodology to determine the subset of heuristics for hyperheuristics.

```
(1) s0 = GenerateInitialSolution
(2) s* = ImprovementStage(s0)
(3) while !StopCondition() do
(4)     s' = SimpleRandomPerturbation(s*)
(5)     s*' = ImprovementStage(s')
(6)     if f(s*') < f(s*) then
(7)         s* = s*'
(8)     end if
(9) end while
(10) return s*
```

Algorithm 1: High-level iterated local search (ILS).

us to modify the methodology solution. We used the same hyperheuristic framework, and according to each class, we gave a different pool of low-level heuristics. For example, if for class 1 the best heuristics were H_1, H_8, H_{10}, and H_5, these were our pool for the hyperheuristic. With a similar procedure, the design was done for each class. After this process, we trained each hyperheuristic on their respective classes for the next stage.

7. Experimental Results

This section describes our experiments in detail for graph coloring and CVRP benchmarks used in this paper. We give the configuration for the implementation of the iterated local search hyperheuristic. Finally, we described the statistical tests that we used to compare our results with the experimental methodology.

Our approach was implemented in JAVA language with JDK 1.8 using the IDE NetBeans IDE 8.2. The experiments were executed on a computer with processor Intel i7-7700U, 2.6 GHz, 16 GB DDR3 RAM, and operating system Windows 10 Home. The tests presented in this work were executed in a common notebook, with a single processor; it is showing the effectiveness of the exposed methodology.

For each heuristic, a limit of 100,000 function calls was given in each test run for all instances. We applied the Shapiro–Wilks test to check if the data results were normal or not, hence choosing a better representative (average or medium). If the data behavior is according to a normal distribution, the average was taken as representative and otherwise the median.

7.1. Heuristics Results for Graph Coloring and CVRP

7.1.1. Graph Coloring. We used the benchmark proposed for the second DIMACS challenge on graph coloring [48] and this is tested with 41 runs. In Tables 5 and 6, we show our results. We denote the best results with a bold face, and only *the myciel2* instance was solved with the application of the individual heuristics in their optimum.

We applied a nonparametric test to verify that there are differences between the performance of each heuristic. Table 7 shows the ranges obtained in the three statistical tests for graph coloring instances. The three omnibus tests indicated there are significant differences between the heuristics. The heuristics which obtained the highest ranks in the three tests were H_6 and H_{11}, i.e., these have the worst performance. H_3 and H_8 were the best results for the problems, and this is because in all omnibus tests those heuristics obtained the lowest ranks (see Table 7 marked in bold). The time of each run is reported in Table 8 in.

```
(1) ls ← IncumbentSolution
(2) while! LocalStopCriteria() do
(3)     hi = Perturbate()
(4)     ls* = apply(hi, ls)
(5)     if f(ls*) < f(ls) then
(6)         ls = ls*'
(7)     end if
(8) end while
(9) return ls
```

Algorithm 2: Improvement stage.

Table 5: Heuristics results for graph coloring instances.

#	Name	Nodes	Colors	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
1	1-Insertions_4	67	3	12	18	**11**	15	10	66	61	13	25	28	75
2	2-Insertions_3	37	3	4	7	**4**	5	4	24	15	4	11	10	29
3	2-Insertions_4	149	3	33	41	**16**	32	32	148	158	29	45	47	155
4	3-Insertions_3	56	3	5	9	**4**	4	5	33	24	4	9	9	40
5	4-Insertions_3	79	2	14	19	16	13	**11**	61	66	13	25	22	71
6	anna	138	10	17	17	14	17	16	59	45	**13**	35	30	68
7	asH_3 31GPIA	662	3	99	**76**	85	96	82	1067	724	89	164	156	1082
8	asH_6 08GPIA	1216	3	1497	1448	**729**	1541	1370	2003	1120	1083	1513	1485	2066
9	asH_9 58GPIA	1916	3	2618	2553	**1900**	2621	2462	3161	1999	2039	2616	2624	3206
10	david	87	10	16	18	**14**	16	16	49	48	15	35	37	58
11	DSJC1000.1	1000	19	1847	1801	**610**	1837	1781	2565	1019	942	1789	1821	2634
12	DSJC1000.5	1000	82	2432	2513	**838**	2425	2379	3130	1126	1230	2327	2327	3234
13	DSJC1000.9	1000	214	1856	2073	**828**	1850	1824	2354	1204	1062	2375	1768	2414
14	DSJC125.1	125	4	48	59	**36**	48	54	161	97	37	56	66	173
15	DSJC125.5	125	15	113	156	**72**	114	124	275	133	68	250	112	298
16	DSJC125.9	125	42	125	173	**77**	124	131	210	132	72	208	112	221
17	DSJC250.5	250	25	350	415	**146**	349	368	658	278	157	339	326	681
18	DSJC250.9	250	70	311	388	**152**	315	319	477	234	153	487	294	485
19	DSJC500.1	500	11	629	646	**197**	634	598	1085	532	279	623	614	1114
20	DSJC500.5	500	47	927	1007	**290**	911	905	1381	471	368	859	853	1412
21	DSJC500.9	500	122	771	927	**326**	759	763	1044	487	382	1051	731	1129
22	DSJR500.1	500	11	124	130	**33**	125	119	334	128	46	129	127	343
23	DSJR500.1c	500	84	1257	1439	**203**	1262	1259	1566	1162	475	1539	1210	1589
24	DSJR500.5	500	121	376	425	**201**	372	369	619	299	211	627	352	662
25	fpsol2.i.1	496	64	141	183	**82**	139	122	263	169	91	255	262	267
26	fpsol2.i.2	451	29	52	81	59	**51**	52	337	281	52	204	215	348
27	fpsol2.i.3	425	29	51	82	55	**51**	54	364	272	54	211	232	368
28	games120	120	8	16	24	11	17	17	82	34	**10**	23	23	88
29	homer	561	12	23	**22**	22	24	23	165	139	24	53	44	176
30	huck	74	10	13	13	**12**	13	13	46	38	12	27	24	47
31	inithx.i.1	864	53	**78**	95	81	80	80	423	320	115	264	263	444
32	inithx.i.2	645	30	55	100	62	**56**	59	525	459	85	350	324	531
33	inithx.i.3	621	30	59	108	65	**56**	63	510	462	78	338	331	539
34	jean	80	9	13	**12**	13	14	12	47	31	13	27	26	46
35	le450_15a	450	14	292	293	**103**	291	291	580	410	137	292	289	597
36	le450_15b	450	14	289	297	**106**	290	285	589	427	125	300	288	624
37	le450_15c	450	14	716	762	**358**	713	713	1187	887	424	734	707	1179
38	le450_15d	450	14	723	751	**371**	718	733	1163	799	418	728	725	1197
39	le450_25a	450	24	160	162	**50**	158	153	372	229	63	164	162	395
40	le450_25b	450	24	156	161	**51**	164	160	377	188	58	156	156	404
41	le450_25c	450	24	407	417	**150**	400	393	736	510	180	396	394	780
42	le450_25d	450	24	400	422	**150**	400	403	736	461	179	397	394	791
43	le450_5a	450	4	780	772	**484**	755	738	1195	966	551	753	771	1209
44	le450_5b	450	4	769	768	**498**	759	738	1192	886	565	764	769	1280

Table 6: Heuristics results for graph coloring instances.

#	Name	Nodes	Colors	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
45	le450_5c	450	4	1459	1476	**909**	1422	1394	2024	1692	1083	1428	1464	2073
46	le450_5d	450	4	1464	1442	**843**	1449	1385	1981	1712	1060	1431	1436	2004
47	miles1000	128	41	52	58	**49**	51	49	133	75	49	129	70	137
48	miles1500	128	73	81	107	82	82	**81**	149	94	79	123	100	169
49	miles250	128	7	14	14	15	15	**13**	73	37	16	24	28	75
50	miles500	128	19	26	24	25	25	**24**	87	48	24	81	43	90
51	miles750	128	30	38	41	38	38	**37**	102	64	37	101	53	119
52	mugg100_1	100	3	5	12	**4**	6	5	55	19	5	12	12	57
53	mugg100_25	100	3	5	12	**4**	6	6	49	25	4	12	14	55
54	mugg88_1	88	3	5	10	**4**	5	5	43	19	5	9	10	50
55	mugg88_25	88	3	5	10	**4**	5	6	46	20	5	11	11	61
56	mulsol.i.1	197	48	60	68	60	60	**57**	142	76	59	131	80	153
57	mulsol.i.2	188	30	49	56	48	50	**47**	172	118	48	143	95	186
58	mulsol.i.3	184	30	50	56	48	49	**47**	177	117	48	160	99	176
59	mulsol.i.4	185	30	49	53	50	49	**48**	165	119	48	142	94	192
60	mulsol.i.5	186	30	50	62	**48**	50	50	171	114	48	107	100	191
61	myciel3	11	3	4	6	4	**3**	**3**	10	6	4	10	4	12
62	myciel4	23	4	**5**	9	6	**5**	**5**	24	12	5	15	9	27
63	myciel5	47	5	8	13	8	9	**8**	53	32	**8**	20	17	56
64	myciel6	95	6	18	24	**12**	19	19	135	105	13	37	51	143
65	myciel7	191	7	84	96	**18**	85	78	347	303	23	107	136	328
66	qg.order30	900	29	554	544	**152**	554	525	912	223	227	533	531	953
67	qg.order40	1600	39	1174	1105	627	1160	1078	1626	**378**	558	1131	1128	1671
68	qg.order60	3600	59	3076	2907	2425	3079	2897	3692	**831**	2234	3070	3008	3738
69	queen10_10	100	10	41	58	37	38	**35**	155	65	37	60	71	163
70	queen11_11	121	10	57	76	54	57	**50**	209	101	53	87	82	216
71	queen12_12	144	12	51	74	52	54	49	224	93	**48**	81	74	233
72	queen13_13	169	12	70	107	**67**	74	68	281	112	69	108	117	299
73	queen14_14	196	15	58	83	54	56	**53**	288	117	53	93	101	317
74	queen15_15	225	15	75	111	**71**	81	74	370	138	74	124	112	368
75	queen16_16	256	16	85	108	83	90	80	407	159	**79**	125	117	438
76	queen5_5	25	4	17	38	16	17	**15**	45	25	17	25	23	50
77	queen6_6	36	6	19	27	17	17	18	49	28	**16**	47	27	55
78	queen7_7	49	6	29	51	**27**	30	33	90	42	30	36	39	87
79	queen8_12	96	11	29	52	30	32	25	138	60	**28**	48	48	138
80	queen8_8	64	8	32	56	**25**	31	33	100	45	27	80	42	104
81	queen9_9	81	9	31	48	**33**	30	27	124	55	34	49	58	129
82	school1	385	13	997	1008	**626**	1009	987	1423	1330	691	1076	1037	1440
83	school1_nsh	352	13	723	742	**445**	735	720	1100	1048	490	734	781	1135
84	will199GPIA	701	3	438	418	438	453	**430**	1762	1542	442	547	555	1788
85	zeroin.i.1	211	48	62	72	62	61	**60**	147	80	60	126	89	159
86	zeroin.i.2	211	29	42	55	**41**	**41**	**41**	165	113	45	163	89	169
87	zeroin.i.3	206	29	40	55	43	40	**39**	179	111	40	162	91	164

Table 7: Statistic test results for graph coloring instances for Friedman (F), Friedman Aligned (FA), and Quade (Q).

	Statistics	p value	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
F	255.72	0	4.64	6.40	2.18	4.68	3.33	10.01	7.52	2.48	7.40	6.47	10.89
FA	77.86	0	374.88	461.35	209.30	370.94	340.22	820.66	575.50	212.84	568.01	505.86	829.45
Q	99.04	0	5.27	6.49	1.91	5.24	3.68	9.99	6.84	2.44	7.11	6.13	10.90

Table 8: Time for one run of each heuristic in seconds for graph coloring instances.

Instance	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
1-Insertions_4	15	14	18	15	14	11	13	6	15	14	13
2-Insertions_3	14	14	17	14	14	11	12	7	14	14	13
2-Insertions_4	16	16	17	17	16	12	14	7	16	16	14
3-Insertions_3	14	14	17	14	14	11	12	6	14	14	13
4-Insertions_3	15	14	18	15	14	11	15	7	15	15	13

Table 8: Continued.

Instance	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
anna	16	15	15	16	15	12	9	5	16	16	14
asH_3 31GPIA	58	44	31	46	44	43	61	27	46	47	45
asH_6 08GPIA	201	172	81	178	175	158	156	93	160	161	134
asH_9 58GPIA	488	321	140	370	354	371	317	203	371	371	275
david	15	14	14	15	14	11	9	5	15	15	14
DSJC1000.1	67	62	25	64	63	73	24	30	63	68	31
DSJC1000.5	51	53	21	55	59	48	17	23	59	59	21
DSJC1000.9	49	55	21	53	61	48	16	24	73	65	22
DSJC125.1	15	15	17	16	15	12	12	6	15	15	14
DSJC125.5	15	15	15	16	15	12	9	4	15	15	14
DSJC125.9	15	15	13	16	15	12	8	4	16	16	13
DSJC250.5	18	18	14	19	18	14	9	6	19	19	15
DSJC250.9	18	18	14	19	18	14	8	5	20	19	14
DSJC500.1	30	27	17	30	28	24	15	11	29	29	20
DSJC500.5	26	26	15	28	27	22	10	9	29	29	15
DSJC500.9	25	26	15	27	28	22	9	9	33	31	15
DSJR500.1	30	27	17	29	28	25	15	11	29	29	20
DSJR500.1c	26	26	15	28	28	22	9	9	31	30	15
DSJR500.5	25	26	15	27	28	22	9	9	33	31	15
fpsol2.i.1	27	26	15	29	27	24	13	10	30	33	15
fpsol2.i.2	25	25	15	26	25	21	10	9	26	27	16
fpsol2.i.3	24	25	15	25	24	20	10	8	25	26	15
games120	15	15	15	16	15	12	10	5	15	15	14
homer	32	30	18	33	34	27	17	12	32	34	24
huck	14	14	15	15	14	11	9	5	14	14	14
inithx.i.1	52	49	22	56	52	51	19	24	54	71	24
inithx.i.2	33	34	17	35	33	30	15	13	36	37	18
inithx.i.3	32	34	17	34	32	29	14	12	35	36	17
jean	14	15	15	15	14	11	9	5	15	14	14
le450_15a	25	26	16	27	25	22	12	9	27	28	17
le450_15b	25	27	16	27	25	22	12	9	27	28	17
le450_15c	25	27	16	27	25	22	12	9	26	28	17
le450_15d	25	27	16	27	25	22	12	9	27	28	17
le450_25a	24	25	15	26	25	21	10	9	27	27	16
le450_25b	24	25	15	26	25	21	10	9	27	27	15
le450_25c	25	26	15	26	25	21	10	9	27	27	16
le450_25d	25	25	15	27	25	21	10	9	27	27	16
le450_5a	26	31	20	28	26	23	25	12	27	32	22
le450_5b	26	31	20	28	26	23	24	12	27	32	22
le450_5c	25	29	20	27	25	22	24	12	27	28	22
le450_5d	25	29	20	27	25	22	24	12	26	28	22
miles1000	15	15	13	16	15	12	8	41	16	16	13
miles1500	15	16	15	16	15	12	8	41	15	16	13
miles250	15	16	15	16	15	12	10	5	15	16	13
miles500	15	15	14	16	15	12	8	41	16	15	14
miles750	15	15	13	16	15	12	8	4	16	16	13
mugg100_1	15	15	17	15	15	11	13	7	15	15	13
mugg100_25	15	15	17	15	15	11	13	7	15	15	13
mugg88_1	15	15	17	15	15	11	13	6	15	15	13
mugg88_25	15	15	17	15	15	11	13	7	15	15	13
mulsol.i.1	17	18	14	18	17	13	8	5	18	18	14
mulsol.i.2	16	18	14	17	16	13	8	5	17	17	14
mulsol.i.3	16	18	14	17	16	13	8	5	17	17	14
mulsol.i.4	16	18	14	17	16	13	8	5	17	17	14
mulsol.i.5	16	18	14	17	16	13	8	5	17	17	14
myciel3	14	14	17	14	14	10	12	6	14	14	13
myciel4	14	14	16	14	14	10	11	6	14	14	13
myciel5	14	14	16	14	14	10	11	5	14	14	13
myciel6	14	15	15	15	15	11	11	5	15	15	14
myciel7	16	17	16	17	17	13	11	6	17	17	14

Table 8: Continued.

Instance	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
qg.order30	50	53	21	54	51	49	20	23	54	57	26
qg.order40	149	151	46	157	152	187	44	71	159	167	53
qg.order60	1016	1016	184	983	1087	983	191	532	1099	1113	226
queen10_10	15	15	15	15	15	11	9	5	15	15	14
queen11_11	15	15	15	15	15	12	9	5	16	16	14
queen12_12	16	15	14	16	15	12	9	5	16	16	14
queen13_13	16	16	15	17	16	12	9	5	16	16	14
queen14_14	17	17	15	18	17	13	9	5	17	17	15
queen15_15	18	18	15	19	18	14	9	6	18	18	15
queen16_16	18	19	14	19	19	15	9	6	19	19	15
queen5_5	14	14	18	14	14	10	12	6	14	14	14
queen6_6	14	14	16	14	14	10	10	5	14	14	14
queen7_7	14	14	16	14	14	11	11	5	14	14	14
queen8_12	15	15	14	15	14	11	9	5	15	15	14
queen8_8	14	14	16	15	14	11	10	5	14	14	14
queen9_9	14	14	14	15	14	11	10	5	15	15	14
school1	23	24	16	24	23	19	12	8	24	25	17
school1_nsh	21	22	16	22	21	18	11	7	22	23	16
will199GPIA	47	51	35	59	67	51	68	32	54	52	46
zeroin.i.1	17	18	14	18	17	14	8	5	18	18	14
zeroin.i.2	17	18	15	18	17	14	9	5	18	18	14
zeroin.i.3	17	18	15	18	17	13	9	5	18	18	14

7.1.2. Capacitated Vehicle Routing Problem (CVRP). Three sets of the state of the art were used and tested on 41 runs:

(1) Augerat et al. (SET A), 9 instances, proposed in [49]

(2) Christofides, Mingozzi, and Toth (CMT), 14 instances, proposed in [50]

(3) Golden, Wasil, Kelly, and Chao (GWKC), 20 instances, proposed in [51]

(4) Uchoa et al., 9 instances, proposed in [52]

In Table 9, we show the fitness values for the instances and the lowest city cost tour is indicated in bold, where n is the number of nodes, Q is the capacity of each vehicle, and k is the number of vehicles (colors in the case of graph coloring). The time of each run is reported in Table 10.

We applied the same procedure to the statistical tests of Friedman (FT), Alienated Friedman (AFT), and Quade (Qt) to distinguish the behavior of the heuristics set. We established $\alpha = 0.05$ and h_0 as there are no differences between the performance of the heuristics and established h_a as there are differences between the performance of the heuristics. Table 11 shows the ranks obtained in the three statistical tests.

In this case, the heuristic H_5 has the lowest rank for the tests and H_6has the second-lowest rank for QT and FT.

7.2. Selection of Features and Classes by Statistical Tests. According to the steps mentioned in Section 5.1, we must determine first the number of clusters or classes to split all our test instances. In this case, $T = 139$ and $c = 8$.

We considered 8 classes and used k-means clusters and we expected uniformly distributed instances in the clusters. The k-means algorithm was applied with a maximum number of Iterations = 500, initially random starting points. To consider the uniform distribution of classes into clusters, we used the Manhattan distance obtained after the experimental work, with the best results.

Table 12 contains the class details, number of instances per cluster/class, number of GCP (3rd column), or CVRP (4th column) per class, min and max nodes, and min and max number of colors nodes. In this experimentation, clusters 1, 5, 6, and 7 have only GCP instances, clusters 3, 4, and 8 have CVRP instances, and only cluster 2 has both problem domains.

7.3. Training and Test Classifiers for the Instance's Classes. After the heuristic pool design phase for the hyperheuristics, we split our dataset into training and test. The training dataset was created by 125 instances with 15 features (basic + inner) and the unseen instances were made by 15 instances. The results of the classification with Naive Bayes are reported in Table 13.

Table 14 contains the confusion matrix of the process classification. We observed that, for some classes like 3, 4, 7, and 8, the patterns were classified correctly. The rest of the classes have some patterns classified incorrectly, but, e.g., for the 3 patterns of class 1 classified into 5 and 6, we used the same pool of low-level heuristics and this does not represent an issue for the next step.

7.4. Designing and Testing the Hyperheuristic Offline Learning with K-Folds. In the next step, the statistical tests were applied to heuristics and will form the characteristics of our instances, graph coloring, and CVRP per class. In this phase, we choose according to the rankings the heuristics which have $x < (\text{minrank} + \text{maxrank})/2$. If, for example, the Aligned Friedman has a min rank = 776.5 and max rank = 3604.5, the

Table 9: Results for CVRP-capacitated instances.

	Instance	Cities	Vehicles	OP	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
88	CMT1	50	5	524.61	1583	1161	1569	1568	1117	**1141**	1664	1576	1664	1664	1664
89	CMT2	75	10	835.26	2353	**1591**	2351	2349	1728	1898	2577	2547	2577	2577	2577
90	CMT3	100	8	826.14	1904	1821	1939	2017	**1699**	1743	3293	2249	3293	3293	3293
91	CMT4	150	12	1028.42	2578	2674	2520	2519	2347	2462	5096	3115	5096	5096	5096
92	CMT5	199	17	1291.29	3159	4469	6490	3144	3008	3155	6742	6707	6742	6692	6742
93	CMT6	50	6	555.43	1027	1160	1569	1028	941	933	1664	1573	1664	1571	1664
94	CMT7	75	11	909.68	1492	1616	2351	1465	1357	1407	2577	2517	2577	2531	2577
95	CMT8	100	9	865.95	1743	1890	1799	1906	1598	1636	3293	2203	3293	3195	3293
96	CMT9	150	14	1162.55	2308	2768	2489	2325	2185	2256	5096	2971	5096	4918	5096
97	CMT10	199	18	1395.85	4607	4607	2852	2852	2791	2870	6742	6707	6742	6513	6742
98	CMT11	120	7	1042.12	2077	2077	2372	2372	2046	2110	2608	2449	2608	2608	2608
99	CMT12	100	10	819.56	1692	1692	1733	1733	1472	1528	2306	2144	2287	2300	2306
100	CMT13	120	11	1541.14	2040	2040	2041	2041	1827	1906	2608	2476	2608	2577	2608
101	CMT14	100	11	866.37	1717	1717	1609	1609	1444	1517	2306	2171	2306	2270	2306
102	GWKC1	240	9	5623.47	9497	9155	8607	9261	7350	8565	13312	10546	13312	13312	13312
103	GWKC2	320	10	8404.61	14571	16631	15161	13399	12320	10903	20961	16801	20787	20735	21080
104	GWKC3	400	9	11036.2	19197	21734	24161	20070	17167	18793	26117	22304	26117	26117	26117
105	GWKC4	480	10	13590	27046	30204	32382	26643	27133	26504	38732	31793	38892	38892	38892
106	GWKC5	200	5	6460.98	14224	15677	13009	14207	12800	13435	19558	15763	21378	21378	21378
107	GWKC6	280	7	8412.9	15407	21043	14925	15701	14022	14539	26699	17688	26699	26699	26699
108	GWKC7	360	8	10102.7	25909	23841	25909	25909	24611	25718	25909	25909	25909	25909	25909
109	GWKC8	440	10	11635.3	21500	23969	20817	20483	18560	18677	31506	26204	31506	31506	31506
110	GWKC9	255	14	579.71	1102	989	2323	1264	954	916	2581	2319	2581	2581	2581
111	GWKC10	323	16	735.66	1344	1312	3476	1277	1222	1229	3589	3451	3589	3589	3589
112	GWKC11	399	17	912.03	1751	1919	4758	1640	1572	1485	5040	4478	5039	5039	5040
113	GWKC12	483	19	1101.5	2295	1979	6738	2462	2192	2363	5416	6520	6750	6750	6750
114	GWKC13	252	26	857.19	2010	1301	2062	2031	1475	1615	2117	2078	2158	2151	2158
115	GWKC14	320	29	1080.55	2351	1637	2637	2322	2221	2389	2748	2665	2745	2744	2748
116	GWKC15	396	33	1337.87	3262	2062	3322	3207	2793	3025	3366	3308	3366	3364	3366
117	GWKC16	480	36	1611.56	3978	2596	4107	3820	3492	3783	4198	4092	4217	4212	4217
118	GWKC17	240	22	707.76	1754	1518	2163	1729	1611	1669	2376	2165	2376	2376	2376
119	GWKC18	300	27	995.13	2130	2165	3443	2150	2073	2155	3443	3443	3443	3425	3443
120	GWKC19	360	33	1365.6	3238	2808	4540	3024	3226	3575	4713	4466	4713	4713	4713
121	GWKC20	420	38	1817.59	3519	3746	6190	3511	3189	3327	6186	5992	6191	6084	6191
122	A-n32-k5	32	5	784	1538	1476	1431	1474	1265	1255	2230	1662	2229	2229	2230
123	A-n45-k6	45	6	1733.36	1733	1733	1733	1733	1273	1156	1733	1733	1733	1733	1733
124	A-n55-k9	55	9	1073	1660	1413	1453	1642	1439	1328	2376	1901	2376	2376	2376
125	A-n61-k9	61	9	1034	2500	1814	2338	2464	1731	1838	2592	2517	2592	2592	2592
126	A-n62-k8	62	8	1288	1772	1724	2039	1792	1551	1615	3194	2219	3194	3194	3194
127	A-n63-k9	63	9	1314	2480	1879	2452	2396	1813	1835	2677	2564	2677	2677	2677
128	A-n64-k9	65	9	1401	1935	1893	2169	1948	1744	1715	2613	2203	2613	2613	2613
129	A-n65-k9	65	9	1174	2609	1940	2600	2546	1820	1892	2949	2808	2949	2949	2949
130	A-n69-k9	69	9	1159	2149	1803	2080	1979	1626	1687	3220	2477	3220	3220	3220
131	X-n148-k46	148	46	43448	71826	54940	75062	71665	68924	71179	77623	78715	79116	79116	79116
132	X-n153-k22	153	22	21220	36700	39843	52741	36155	34116	34337	71801	56820	71799	71793	71801
133	X-n157-k13	157	13	16876	36364	24507	36364	36364	27917	29210	36364	36364	36364	36364	36364
134	X-n162-k11	162	11	14138	39232	37225	36381	38162	32918	34059	79973	60657	79969	79963	79973
135	X-n367-k17	367	17	22814	71209	85326	97502	72482	68648	70027	154244	112919	154244	154244	154244
136	X-n393-k38	393	39	38260	141432	74629	152513	143398	130939	148728	158058	155203	158058	158058	158058
137	X-n401-k29	401	29	66187	140100	106590	163158	141582	141056	145043	197323	163230	197323	197323	197323
138	X-n411-k19	411	19	19718	67963	79642	81034	67090	63038	66173	173269	88152	173269	173269	173269
139	X-n420-k130	420	130	10798	182376	132745	207213	183053	203757	200726	217483	216871	217483	217483	217483

Table 10: Time for one run of each heuristic in seconds for CVRP-capacitated instances.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
A-n32-k5	7	7	12	7	7	5	7	4	7	7	9
A-n45-k6	7	7	12	7	7	5	6	3	7	7	9
A-n55-k9	7	7	11	7	7	5	5	3	7	7	9
A-n61-k9	7	7	11	7	7	5	5	3	7	7	9
A-n62-k8	7	7	11	8	7	5	6	3	7	7	9
A-n63-k9	7	7	11	7	7	5	5	3	7	7	9
A-n63-k10	7	7	11	7	7	5	5	3	7	7	9
A-n64-k9	7	7	11	7	7	5	5	3	7	7	9
A-n65-k9	7	7	11	7	7	5	5	3	7	7	9
A-n69-k9	7	7	11	8	7	5	5	3	8	7	9
CMT1	7	7	12	7	7	5	7	4	7	7	9
CMT2	7	7	11	7	7	5	5	3	8	8	9
CMT3	7	7	11	8	7	5	6	3	8	8	9
CMT4	8	8	11	8	8	6	6	3	9	9	9
CMT5	8	8	10	9	8	6	5	3	10	10	9
CMT6	7	7	12	7	7	5	6	3	7	7	9
CMT7	7	7	11	8	7	5	5	2	8	7	9
CMT8	7	7	11	8	7	5	6	3	8	8	9
CMT9	8	8	10	8	8	6	5	2	9	9	9
CMT10	8	8	10	9	8	6	5	2	10	10	9
CMT11	7	7	11	8	7	6	6	3	8	8	9
CMT12	7	7	12	7	7	5	7	3	7	7	9
CMT13	7	7	12	7	7	5	6	3	7	7	9
CMT14	7	7	12	7	7	5	7	4	7	7	9
X-n420-k130	11	11	16	11	11	9	11	8	11	11	13
X-n411-k19	11	11	16	11	11	9	11	8	11	11	13
X-n401-k29	11	11	16	11	11	9	10	7	11	11	13
GWKC1	8	8	11	9	8	6	6	3	9	9	9
GWKC2	8	8	11	10	8	6	6	3	10	10	10
GWKC3	9	9	11	11	9	7	8	4	12	12	13
GWKC4	10	10	13	12	10	8	8	3	12	12	11
GWKC5	8	8	13	9	8	6	9	4	9	8	9
GWKC6	8	8	12	9	8	6	7	3	9	9	10
GWKC7	9	9	11	9	9	7	7	3	10	10	10
GWKC8	9	9	11	11	9	7	7	3	12	12	10
GWKC9	8	8	11	9	8	6	5	3	10	10	10
GWKC10	9	9	11	10	9	7	6	3	12	11	10
GWKC11	9	9	11	11	9	7	6	3	13	13	10
GWKC12	10	10	11	12	10	8	6	3	15	15	11
GWKC13	8	8	11	10	9	7	5	2	12	12	10
GWKC14	9	9	11	10	9	7	5	3	14	14	10
GWKC15	10	10	11	11	9	7	5	3	16	16	10
GWKC16	10	10	11	12	10	8	5	3	19	18	10
GWKC17	8	8	10	9	8	6	5	2	11	11	10
GWKC18	9	9	10	10	9	7	5	2	13	13	10
GWKC19	9	9	11	11	9	7	5	3	16	15	10
GWKC20	10	10	11	12	10	8	5	3	18	18	11
X-n148-k46	11	11	16	11	11	9	11	8	11	11	13
X-n153-k22	11	11	16	11	11	9	11	7	11	11	13
X-n157-k13	11	11	16	11	11	9	10	7	11	11	13
X-n162-k11	11	11	15	11	11	9	10	7	11	11	13
X-n367-k17	11	11	15	11	11	9	10	7	11	11	13
X-n393-k38	11	11	15	11	11	9	10	7	11	11	13

Table 11: Statistic test results for capacitated vehicle routing problem instances for Friedman (F), Friedman Aligned (FA), and Quade (Q).

	Statistics	p value	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
F	420.84	0	4.49	8.79	9.69	3.53	5.59	4.16	1.60	2.46	9.32	6.92	9.45
FA	47.3	0	170.82	441.25	443.96	143.53	280.22	169.32	127.19	147.94	441.70	342.00	443.57
Q	79.97	0	4.29	8.87	9.72	3.68	5.59	3.96	1.67	2.57	9.26	6.91	9.47

Table 12: Summary cluster/classes with k–means algorithm.

Cluster	# of instances	%	GC	CVRP	Min nodes	Max nodes	Min colors/vehicles	Max colors/vehicles
1	18	12.9	18	0	11	149	3	6
2	10	7.2	8	2	420	3600	3	214
3	11	7.9	0	11	148	480	5	46
4	5	3.6	0	5	367	420	17	130
5	19	13.7	19	0	64	256	7	30
6	31	22.3	31	0	125	900	11	122
7	11	7.9	11	0	352	701	3	14
8	34	24.5	0	34	32	483	5	33

Table 13: Classification value results.

Correctly classified	114	91.20%
Incorrectly classified	11	8.80%
TP rate	0.912	
FP rate	0.022	
Precision	0.918	
Recall	0.912	

Table 14: Confusion matrix, TP rate (TPR), FP rate (FPR), and precision (P) for each class.

Class	1	2	3	4	5	6	7	8	TPR	FPR	P
1	13	0	0	0	2	1	0	0	0.813	0	1
2	0	7	0	0	0	0	0	2	0.778	0	1
3	0	0	10	0	0	0	0	0	1	0	1
4	0	0	0	4	0	0	0	0	1	0	1
5	0	0	0	0	12	5	0	0	0.706	0.019	0.857
6	0	0	0	0	0	27	1	0	0.964	0.062	0.818
7	0	0	0	0	0	0	10	0	1	0.009	0.909
8	0	0	0	0	0	0	0	31	1	0.021	0.939

limit value for considering a heuristic must be 2190.5. Because heuristic 11 has the worst performance for both datasets, we did not consider this heuristic for this experimentation phase.

Table 15 shows the ranks and the heuristic fitness for each class, e.g., class 1 for QT, and AFT has the same heuristics H_5, H_6, H_4, H_1, H_2, and H_3, while FT does not consider H_3. We only consider H_5, H_6, H_4, H_1, H_2, and H_3 as a minimum set.

The selected heuristics for each class can be summarized into five groups:

(1) H_3, H_8, H_5, H_4, and H_1
(2) H_3, H_8, H_7, H_5, and H_2
(3) H_5, H_6, H_2, H_4, H_1, and H_3
(4) H_2, H_1, H_5, H_4, and H_6
(5) H_3, H_8, H_5, and H_4

Therefore, we design 5 algorithms for these 8 classes. This means the hyperheuristic with the high-level search strategy was the same, but the heuristics pool set was different according to these 5 subset heuristics. We trained and tested the hyperheuristic on short-length pools. We left 10% of instances as unseen for the hyperheuristics and the results are shown in Table 16.

The hyperheuristic configuration was 10 iterations for local search and 100,000 function calls. For some GCP instances, we got the optimum number of colors (denoted in bold in Table 16). Besides, for the instances A-n45-k6, A-n55-k9, A-n62-k8, A-n64-k9, CMT13, GWKC1, GWKC2, GWKC3, and X-n148-k46, we get values near to the optimal with a maximum of 20% of the distance.

7.5. Classification of the Test Instances and Application of the Hyperheuristic to the Corresponding Instance. Finally, for the 14 unseen instances, we used the Naive Bayes classifier, and it determines the class for these instances. Later, we applied the hyperheuristic with the corresponding pool heuristics according to the previous design and we obtained the results shown in Tables 17 and 18.

Table 17 shows the confusion matrix, TP rate, FP rate, and precision of the classification test. In these results, two patterns that belong to class five were classified incorrectly, but this does not affect the hyperheuristic solution because this class shares the same heuristics with class 6.

7.6. Statistical Comparison of Results. Finally, to compare if there are differences between the results applying the methodology and without applying the methodology, an experiment was carried out where the hyperheuristics were executed 33 times with the entire set of heuristics and 100,000 function calls. The results are shown in Table 19.

First, the statistical distributions followed by each set of results by class were analyzed, that is, the Shapiro–Wilks test [53] was applied to determine if the results of the methodology, hyperheuristic without methodology (HHPC), and the optimal state of the art followed a normal distribution.

Table 15: Heuristics ranks per classes, Friedman (Fr), Aligned Friedman (A-Fr), and Quade (Qu).

	Class 1						Class 2					
*Fr	H_3 1.50	H_8 1.69	H_5 3.78	H_1 4.03	H_4 4.08	H_2 6.69	H_3 2.8	H_8 3	H_7 3.4	H_5 4	H_2 5.7	H_{10} 6.25
*A-Fr	H_3 627	H_8 638	H_1 1089	H_5 1091	H_4 1102	H_2 1760.5	H_3 293	H_8 320	H_7 335	H_5 428	H_2 508	H_4 540
*Qu	H_3 1.55	H_8 1.59	H_5 3.77	H_4 4.04	H_1 4.15	H_2 6.81	H_3 2.24	H_8 2.73	H_7 2.98	H_5 4.18	H_2 5.51	H_{10} 6.39
	Class 3						Class 4					
*Fr	H_5 1.64	H_6 2.18	H_2 4.27	H_4 4.27	H_1 4.82	H_3 5.18	H_5 2.4	H_2 2.6	H_1 2.8	H_4 3.6	H_6 3.6	H_3 6
*A-Fr	H_5 242	H_6 283	H_4 403	H_2 418	H_1 419	H_3 568	H_2 33	H_1 74	H_5 78	H_4 79	H_6 91	H_3 128
*Qu	H_5 1.71	H_6 2.26	H_2 4.20	H_4 4.52	H_1 4.88	H_3 5.38	H_2 2.07	H_1 2.67	H_5 2.67	H_4 3.67	H_6 3.93	H_3 6
	Class 5						Class 6					
*Fr	H_8 1.45	H_3 1.55	H_5 3.53	H_1 4.37	H_4 4.68	H_2 6.03	H_3 1.32	H_8 1.69	H_5 4.35	H_4 4.92	H_1 5.05	H_{10} 5.92
*A-Fr	H_8 646	H_3 651	H_5 1144.5	H_1 1197	H_4 1225.5	H_2 2010.5	H_3 1306	H_8 1345	H_5 4112	H_4 4211	H_1 4227.50	H_2 5058
*Qu	H_8 1.47	H_3 1.53	H_5 3.67	H_1 4.44	H_4 4.69	H_2 5.78	H_3 1.48	H_8 1.53	H_5 3.90	H_4 4.55	H_1 4.60	H_2 6.47
	Class 7						Class 8					
*Fr	H_3 1.32	H_8 2.45	H_5 3.45	H_4 5.32	H_1 5.86	H_9 5.91	H_5 1.47	H_6 2.38	H_2 3.44	H_4 4.21	H_1 4.59	H_3 5.46
*A-Fr	H_3 108.50	H_8 171	H_5 477.50	H_4 578.50	H_1 637.50	H_9 642	H_5 2328	H_6 2691	H_2 2989	H_4 3847	H_1 4070	H_3 6169.50
*Qu	H_3 1.12	H_8 2.17	H_5 3.65	H_4 5.09	H_9 5.74	H_1 5.82	H_5 1.50	H_6 2.28	H_2 3.27	H_4 4.46	H_1 4.70	H_3 5.43

Table 16: Hyperheuristics results for graph coloring and CVRP.

#	C	Median	DE	Average	Opt	#	C	Median	DE	Average	Opt	#	C	Median	DE	Average	Opt
4	1	**3**	2	4	3	30	5	**10**	4	12	10	7	7	22	284	162	3
1	1	4	39	30	3	34	5	10	5	12	9	19	7	160	212	490	11
3	1	4	12	9	3	49	5	8	10	15	7	37	7	326	185	623	14
4	1	**3**	5	5	3	50	5	20	11	26	19	38	7	321	183	620	14
5	1	**3**	10	7	3	51	5	32	14	40	30	43	7	407	160	674	4
14	1	18	28	40	4	65	5	**7**	73	72	7	44	7	402	168	672	4
53	1	**3**	8	6	3	69	5	22	25	36	10	45	7	294	398	1,147	4
54	1	**3**	5	5	3	70	5	37	33	56	10	46	7	293	390	1,156	4
55	1	**3**	6	5	3	71	5	32	42	55	12	83	7	340	171	658	13
61	1	**3**	0	3	3	72	5	52	55	79	12	84	7	320	413	548	3
62	1	**4**	2	5	4	73	5	35	58	64	15	123	8	1,821	261	1,524	1,733
63	1	**5**	6	7	5	75	5	60	80	101	16	124	8	1,147	517	1,598	1,073
64	1	**6**	24	19	6	80	5	15	15	25	8	125	8	1,366	468	2,082	1,034
76	1	**4**	7	8	4	81	5	18	21	29	9	126	8	1,410	767	2,052	1,288
77	1	8	8	12	6	16	6	60	33	99	42	127	8	1,582	437	2,145	1,314
78	1	16	11	23	6	18	6	127	82	255	70	128	8	1,510	455	1,975	1,401
8	2	573	365	1,176	3	20	6	247	285	715	47	129	8	1,433	583	2,271	1,174
9	2	1,011	510	2,190	3	21	6	285	206	628	122	130	8	1,420	747	2,120	1,159
11	2	541	535	1,435	19	22	6	21	59	96	11	88	8	909	291	1,339	525
13	2	771	478	1,542	214	24	6	181	97	318	121	97	8	2,469	1,897	4,622	1,396
23	2	153	441	962	84	25	6	75	43	133	64	98	8	1,707	342	2,174	1,042
117	2	2,179	776	3,343	1,612	26	6	42	77	98	29	99	8	1,116	426	1,605	820
121	2	2,858	1,520	4,528	1,818	27	6	40	80	100	29	100	8	1,548	418	2,051	1,541
67	2	318	330	897	39	29	6	14	35	35	12	101	8	1,117	420	1,564	866
68	2	753	689	2,522	59	31	6	68	86	131	53	90	8	1,441	787	2,105	826
103	3	9,429	4,587	13,967	8,405	32	6	45	130	142	30	91	8	2,049	1,313	3,099	1,028
104	3	13,211	5,617	19,240	11,036	33	6	46	136	145	30	92	8	2,619	1,842	4,702	1,291
106	3	11,376	4,025	14,762	6,461	35	6	85	104	239	14	93	8	815	342	1,165	555
107	3	12,566	6,118	17,696	8,413	36	6	85	108	238	14	94	8	1,188	572	1,800	910

Table 16: Continued.

#	C	Median	DE	Average	Opt	#	C	Median	DE	Average	Opt	#	C	Median	DE	Average	Opt
108	3	13,221	4,905	23,107	10,103	39	6	37	68	126	24	95	8	1,436	727	2,024	866
109	3	14,390	7,055	21,411	11,635	40	6	35	69	124	24	102	8	6,023	3,021	8,874	5,624
131	3	50,634	9,126	68,656	43,448	42	6	127	137	332	24	111	8	1,056	1,172	2,201	736
132	3	30,323	17,173	47,025	21,220	47	6	43	15	53	41	112	8	1,340	1,671	2,996	912
133	3	22,175	5,420	32,883	16,876	48	6	74	15	86	73	113	8	1,744	2,287	4,038	1,102
134	3	28,769	21,922	45,678	14,138	56	6	50	17	62	48	114	8	1,138	422	1,658	857
135	4	65,098	59,475	160,617	22,814	57	6	40	28	59	30	115	8	1,447	522	2,154	1,081
136	4	67,184	29,544	127,827	38,260	58	6	37	26	57	30	116	8	1,811	574	2,737	1,338
137	4	97,074	34,915	149,750	66,187	60	6	40	27	58	30	118	8	1,300	462	1,810	708
139	4	125,875	27,756	192,093	10,798	66	6	131	193	417	29	119	8	1,623	837	2,587	995
10	5	**10**	6	15	10	85	6	53	17	65	48	120	8	2,370	1,032	3,519	1,366
15	5	45	48	95	15	86	6	34	28	55	29	110	8	823	803	1,603	580
28	5	**8**	14	14	8	87	6	33	26	53	29						

The best results are highlighted in bold.

Table 17: Confusion matrix, TP rate (TPR), FP rate (FPR), and precision (P) for each test instance.

Class	1	2	3	4	5	6	7	8	TPR	FPR	P
1	2	0	0	0	0	0	0	0	1	0	1
2	0	1	0	0	0	0	0	0	1	0	1
3	0	0	1	0	0	0	0	0	1	0	1
4	0	0	0	1	0	0	0	0	1	0	1
5	0	0	0	0	0	2	0	0	0	0	—
6	0	0	0	0	0	3	0	0	1	0.182	0.6
7	0	0	0	0	0	0	1	0	1	0	1
8	0	0	0	0	0	0	0	3	1	0	1

Table 18: Fitness results of test instances (hyperheuristic).

Instance	Class	Fitness	Instance	Class	Fitness
anna	1	3	DSJC250.5	6	27
mugg100_1	1	3	le450_25c	6	26
DSJC1000.5	2	83	mulsol.i.4	6	30
GWKC4	3	14255.68	school1	7	13
X-n411-k19	4	20005.37	A-n32-k5	8	876.19
queen15_15	5	29	CMT2	8	911.37
queen8_12	5	21	CMT9	8	1258.57

The test was applied with a $\propto = 0.05$. The results of the test are shown in Table 20, and the data shown in Tables 16, 18, and 19 were taken for the tests. It should be noted that the results of the methodology only for clusters 4 and 5 were normal, the results of the optimal state of the art only for only cluster 4 were normal, and the results of HHPC for clusters 1, 4, 5, and 7 were normal.

Student' *t*-tests for methodology and state of the art were applied. For the state of the art, methodology, and HHPC, we established $\propto = 0.05$ as a level of significance. The null and alternative hypotheses for methodology and HHPC are as follows:

(i) h_0: there are no differences between the performance of hyperheuristic with the methodology and without the methodology

(ii) h_a: there are differences between the performance of hyperheuristics with the methodology and without the methodology

For methodology and state of the art,

(i) h_0: there are no differences between the performance of hyperheuristics and the optimal state of the art

(ii) h_a: there are differences between the performance of hyperheuristics

The statistical results of the tests are shown in Table 21. With these values, we can observe the following:

(i) *Methodology and HHPC*. It can be inferred that the results of the methodology are significantly different from those of hyperheuristics with the whole set of heuristics. This means that the methodology improved performance and allowed limiting the set of heuristics for each of the clusters.

(ii) *Methodology and State of the Art*. It can be inferred that no statistical evidence was found that the results of the methodology differ from the optimal ones of the state of the art, except in clusters 5 and 7. This is because it is where there are more atypical data or that they were badly classified which opens an area of opportunity for the refinement of the methodology.

Table 19: Results of hyperheuristic with the complete set heuristics and without methodology (HHPC), cluster (C), and T which are the training set instances.

Instance	c	HHPC	Instance	c	HHPC	Instance	c	HHPC	Instance	c	HHPC
1-Insertions_4	1	34	X-n367-k17	4	155287	le450_15b	6	242	CMT12	8	2345
2-Insertions_3	1	11	X-n393-k38	4	158696	le450_25a	6	137	CMT13	8	2769
2-Insertions_4	1	60	X-n401-k29	4	198461	le450_25b	6	140	CMT14	8	2361
3-Insertions_3	1	14	X-n420-k130	4	218854	le450_25d	6	341	CMT3	8	3335
4-Insertions_3	1	28	david	5	27	miles1000	6	72	CMT4	8	5185
DSJC125.1	1	70	DSJC125.5	5	138	miles1500	6	97	CMT5	8	6785
mugg100_25	1	15	games120	5	31	mulsol.i.1	6	74	CMT6	8	1671
mugg88_1	1	20	huck	5	22	mulsol.i.2	6	87	CMT7	8	2582
mugg88_25	1	14	jean	5	20	mulsol.i.3	6	74	CMT8	8	3355
myciel3	1	6	miles250	5	24	mulsol.i.5	6	70	GWKC1	8	13602
myciel4	1	9	miles500	5	41	qg.order30	6	351	GWKC10	8	3630
myciel5	1	26	miles750	5	51	zeroin.i.1	6	78	GWKC11	8	5057
myciel6	1	48	myciel7	5	117	zeroin.i.2	6	70	GWKC12	8	6775
queen5_5	1	26	queen10_10	5	82	zeroin.i.3	6	78	GWKC13	8	2191
queen6_6	1	29	queen11_11	5	108	asH_3 31GPIA	7	433	GWKC14	8	2793
queen7_7	1	42	queen12_12	5	105	DSJC500.1	7	477	GWKC15	8	3413
asH_6 08GPIA	2	746	queen13_13	5	141	le450_15c	7	618	GWKC17	8	2398
asH_9 58GPIA	2	1157	queen14_14	5	143	le450_15d	7	643	GWKC18	8	3467
DSJC1000.1	2	1090	queen16_16	5	184	le450_5a	7	632	GWKC19	8	4756
DSJC1000.9	2	1625	queen8_8	5	50	le450_5b	7	699	GWKC9	8	2604
DSJR500.1c	2	930	queen9_9	5	69	le450_5c	7	861	A-n32-k5	T	2235
GWKC16	2	4264	DSJC125.9	6	119	le450_5d	7	792	anna	T	27
GWKC20	2	6248	DSJC250.9	6	255	school1_nsh	7	671	CMT2	T	2611
qg.order40	2	623	DSJC500.5	6	656	will199GPIA	7	898	CMT9	T	5145
qg.order60	2	2098	DSJC500.9	6	611	A-n45-k6	8	1856	DSJC1000.5	T	1765
GWKC2	3	21504	DSJR500.1	6	100	A-n55-k9	8	2474	DSJC250.5	T	336
GWKC3	3	26716	DSJR500.5	6	314	A-n61-k9	8	2681	GWKC4	T	39592
GWKC5	3	21807	fpsol2.i.1	6	137	A-n62-k8	8	3200	le450_25c	T	321
GWKC6	3	27213	fpsol2.i.2	6	136	A-n63-k9	8	2774	mugg100_1	T	17
GWKC7	3	26506	fpsol2.i.3	6	131	A-n64-k9	8	2709	mulsol.i.4	T	104
GWKC8	3	32073	homer	6	50	A-n65-k9	8	2977	queen15_15	T	164
X-n148-k46	3	80822	inithx.i.1	6	190	A-n69-k9	8	3258	queen8_12	T	52
X-n153-k22	3	72255	inithx.i.2	6	188	CMT1	8	1709	school1	T	850
X-n157-k13	3	37747	inithx.i.3	6	178	CMT10	8	6763	X-n411-k19	T	174414
X-n162-k11	3	80314	le450_15a	6	251	CMT11	8	2624			

Table 20: Normality test results of the methodology, state of the art, and hyperheuristics with complete pool.

Cluster	Methodology	State of the art	HHPC
1	$2.26E-05$	$2.69E-04$	$1.13E-01$
2	0.0309	$1.57E-04$	$7.52E-03$
3	0.01438	$1.80E-03$	$4.28E-03$
4	0.3551	$7.91E-01$	$3.29E-01$
5	0.05466	$1.05E-03$	$9.61E-02$
6	$1.20E-05$	$6.53E-05$	$2.25E-05$
7	0.0255	$3.32E-03$	$7.31E-01$
8	$4.95E-08$	$2.40E-09$	$4.59E-07$

Table 21: Results of the Student t-tests for methodology and Edo. Art and methodology and HHPC.

Cluster	Methodology and Edo. Art	Methodology and HHPC
	p_{value}	p_{value}
1	0.1641	5.12E-05
2	0.1489	0.1496
3	0.3211	0.02586
4	0.02814	0.004389
5	0.00601	0.0002388
6	0.01144	0.0009709
7	4.43E-07	6.65E-06
8	0.06235	2.52E-05

8. Conclusion

In this work, a methodology is proposed to select low-level heuristics for a hyperheuristic approach to offline learning oriented to the solution of instances of different constraint satisfaction problems. The proposal was applied to two different problems well known and studied in the state of the art, which were the coloring of graphs and the vehicle routing problem with a specific capacity, GCP and CVRP, respectively.

The methodology is focused on optimizing the number of heuristics that can be applied to different constraint satisfaction problems in a hyperheuristic approach. Information on the performance of an original set of heuristics for the instances of the problem is obtained from the different problems. The performance information is used to generate characteristic vectors for each instance, which is used to generate equivalence classes of instances of the problem. The grouping in classes allows to identify the heuristics that apply to each class and from that information, a reduction of the number of heuristics necessary to obtain good solutions in the instance of each class is made and to reduce the total number of heuristics that can be applied in the hyperheuristic approach to solving the problems involved.

In the application to the GCP and CVRP, the information on the performance of the heuristics was obtained through a metalearning process, and this information was used to obtain the basic and internal characteristics of the instances. The instances were grouped into 5 classes using the k-means algorithm with the Mahalanobis metric. For each class, the sets of heuristics that could be applied to all their instances were identified, and through a process of hierarchization and cutoff criteria, the number of heuristics per class was reduced.

For training and testing, the Naive Bayes classifier and information on the characteristics of the instances were used. The experimental results show that the hyperheuristic in each class could efficiently solve each instance, and the classifier was able to predict the class for each problem instance.

The identification and reduction of heuristics to find the solution of complex problems is an optimization strategy that can do the search for solutions to problems of satisfaction of restrictions efficiently. The methodology presented allows generating a framework with a level of generality that can be trained to solve different problems of satisfaction of constraints simultaneously under the hyperheuristic approach. Once trained, it can allow finding good solutions to different problems with a common base of heuristics for instances of problems grouped by the efficiency of the solution heuristics.

Finally, the methodology makes it possible to improve the search for solutions to sets of problems by exploring the diversification of some of its components such as classification algorithms, metrics, heuristics, and selection criteria, which may be different for sets of different problems. A study of these possibilities is proposed as future work.

Data Availability

The instances data used to support the findings of this study have been deposited in the graph coloring repository http://vrp.atd-lab.inf.puc-rio.br/index.php/en/, http://archive.dimacs.rutgers.edu/pub/challenge/graph/benchmarks/color/, https://neo.lcc.uma.es/vrp/vrp-instances/capacitated-vrp-instances/. They are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank Tecnológico Nacional de México/I. T. León and Universidad de Guanajuato. This work was supported by the National Council of Science and Technology of Mexico (CONACYT) via the Scholarship for Postgraduate Study 446106 (L. Ortiz) and Research Grant: CÁTEDRAS-2598 (A. Rojas). The authors thank the participation of Valentin Calzada-Ledesma from Instituto Tecnológico Superior de Purísima for the revision and correction of this article.

References

[1] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[2] N. Pillay, "A review of hyper-heuristics for educational timetabling," *Annals of Operations Research*, vol. 239, no. 1, pp. 3–38, 2016.

[3] W. B. Yates and E. C. Keedwell, "An analysis of heuristic subsequences for offline hyper-heuristic learning," *Journal of Heuristics*, vol. 25, no. 3, pp. 399–430, 2019.

[4] E. K. Burke, M. R. Hyde, G. Kendall et al., "A classification of hyper-heuristic approaches: revisited," in *Handbook of Metaheuristics*, pp. 453–477, Springer International Publishing, Berlin, Germany, 2018.

[5] J. Kanda, A. D. Carvalho, E. Hruschka, C. Soares, and P. Brazdil, "Meta-learning to select the best meta-heuristic for the traveling salesman problem: a comparison of meta-features," *Neurocomputing*, vol. 205, pp. 393–406, 2016.

[6] J. R. Rice, "The algorithm selection problem," in *Advances in Computers*, pp. 65–118, Elsevier, Amsterdam, Netherlands, 1976.

[7] I. Amaya, J. C. Ortiz-Bayliss, S. Conant-Pablos, and H. Terashima-Marin, "Hyper-heuristics reversed: learning to combine solvers by evolving instances," in *Proceedings of the 2019 IEEE Congress on Evolutionary Computation (CEC)*, December 2019.

[8] H. Song, I. Triguero, and E. Özcan, "A review on the self and dual interactions between machine learning and optimisation," *Progress in Artificial Intelligence*, vol. 8, no. 2, pp. 143–165, 2019.

[9] J. A. Montesino-Guerra, H. Puga, J. M. Carpio, M. Ornelas-Rodríguez, A. Rojas-Domínguez, and L. Ortiz-Aguilar, "Combinatorial designs on constraint satisfaction problem (VRP)," in *Intuitionistic and Type-2 Fuzzy Logic*

Enhancements in Neural and Optimization Algorithms: Theory and Applications, pp. 509–526, Springer, Cham, Switzerland, 2020.

[10] L. D. M. Ortiz-Aguilar, M. Carpio, H. Puga et al., "Increase methodology of design of course timetabling problem for students, classrooms, and teachers," in *Nature-Inspired Design of Hybrid Intelligent Systems*, pp. 713–728, Springer, Berlin, Germany, 2017.

[11] S. Lin and B. W. Kernighan, "An effective heuristic algorithm for the traveling-salesman problem," *Operations Research*, vol. 21, no. 2, pp. 498–516, 1973.

[12] K. Helsgaun, "General k-opt submoves for the Lin–Kernighan TSP heuristic," *Mathematical Programming Computation*, vol. 1, no. 2-3, pp. 119–163, 2009.

[13] A. Blazinskas and A. Misevicius, "Combining 2-opt, 3-opt and 4-opt with k-swap-kick perturbations for the traveling salesman problem," in *Proceedings of the 17th International Conference on Information and Software Technologies*, pp. 27–29, Kaunas, Lithuania, April 2011.

[14] S. Minton, M. D. Johnston, A. B. Philips et al., "Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems," *Artificial Intelligence*, vol. 58, no. 1–3, pp. 161–205, 1992.

[15] B. S. Baker, "A new proof for the first-fit decreasing bin-packing algorithm," *Journal of Algorithms*, vol. 6, no. 1, pp. 49–70, 1985.

[16] J. Csirik, "The parametric behavior of the first-fit decreasing bin packing algorithm," *Journal of Algorithms*, vol. 15, no. 1, pp. 1–28, 1993.

[17] J. A. Soria-Alcaraz, G. Ochoa, M. Carpio et al., "Effective learning hyper-heuristics for the course timetabling problem," *European Journal of Operational Research*, vol. 238, no. 1, pp. 77–86, 2014.

[18] J. A. Soria-Alcaraz, E. Özcan, J. Swan, G. Kendall, and M. Carpio, "Iterated local search using an add and delete hyper-heuristic for university course timetabling," *Applied Soft Computing*, vol. 40, pp. 581–593, 2016.

[19] W. B. Yates and E. C. Keedwell, "Offline learning for selection hyper-heuristics with Elman networks," in *Lecture Notes in Computer Science*, pp. 217–230, Springer International Publishing, Berlin, Germany, 2018.

[20] J. A. Soria-Alcaraz, G. Ochoa, M. A. Sotelo-Figeroa, and E. K. Burke, "A methodology for determining an effective subset of heuristics in selection hyper-heuristics," *European Journal of Operational Research*, vol. 260, no. 3, pp. 972–983, 2017.

[21] C. Martin, P. Héctor, T. M. Hugo et al., "Methodology of design: a novel generic approach applied to the course timetabling problem," in *Soft Computing Applications in Optimization, Control, and Recognition*, pp. 287–319, Springer, Berlin, Germany, 2013.

[22] P. Brazdil, C. Giraud-Carrier, C. Soares et al., *Metalearning*, Springer, Berlin, Germany, 2009.

[23] A. E. Gutierrez-Rodríguez, S. E. Conant-Pablos, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "Selecting meta-heuristics for solving vehicle routing problems with time windows via meta-learning," *Expert Systems with Applications*, vol. 118, pp. 470–481, 2019.

[24] R. Karp, *Complexity of Computer Computations*, Springer US, Boston, MA, USA, 1972.

[25] A. Kosowski and K. Manuszewski, "Classical coloring of graphs," *Graph Colorings*, vol. 352, pp. 1–19, 2004.

[26] G. B. Dantzig and J. H. Ramser, "The truck dispatching problem," *Management Science*, vol. 6, no. 1, pp. 80–91, 1959.

[27] E. Alba and B. Dorronsoro, "Computing nine new best-so-far solutions for capacitated VRP with a cellular genetic algorithm," *Information Processing Letters*, vol. 98, no. 6, pp. 225–230, 2006.

[28] J. A. Soria-Alcaraz, M. Carpio, P. Hector et al., "Comparison of metaheuristic algorithms with a methodology of design for the evaluation of hard constraints over the course timetabling problem," *Studies in Computational Intelligence*, Vol. 451, Springer, Berlin, Germany, 2013.

[29] M. Carpio, "Modelo integral de asignación optima de carga academica usando un algoritmo heurístico," in *Actas de la Encuentro de Investigación en Ingenieria Electrica*, Zacatecas, Mexico, Febrero 2006.

[30] L. D. M. Ortiz Aguilar, "Diseño de horarios de alumnos y maestros mediante técnicas de soft computing, para una institución educativa," Master's thesis, Instituto Tecnológico de León, León, Mexico, 2016.

[31] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Cambridge, MA, USA, 3rd edition, 2014.

[32] P. Brazdil and C. Giraud-Carrier, "Metalearning and algorithm selection: progress, state of the art and introduction to the 2018 special issue," *Machine Learning*, vol. 107, no. 1, pp. 1–14, 2018.

[33] J. Sakarovitch, "Kleene's theorem revisited," in *International Meeting of Young Computer Scientists*, pp. 39–50, Springer, Berlin, Germany, 1986.

[34] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 160–172, Springer, Berlin, Germany, 2013.

[35] D. W. Scott, "Sturges' rule," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 303–306, 2009.

[36] J. Jokinen, T. Raty, and T. Lintonen, "Clustering structure analysis in time-series data with density-based clusterability measure," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1332–1343, 2019.

[37] X. Xu, J. Li, M. Zhou, J. Xu, and J. Cao, "Accelerated two-stage particle swarm optimization for clustering not-well-separated data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4212–4223, 2020.

[38] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011.

[39] D. de Werra, "Heuristics for graph coloring," in *Computational Graph Theory*, pp. 191–208, Springer, Berlin, Germany, 1990.

[40] H. Derbel, B. Jarboui, and R. Bhiri, "A skewed general variable neighbourhood search algorithm with fixed threshold for the heterogeneous fleet vehicle routing problem," *Annals of Operations Research*, vol. 272, no. 1-2, pp. 243–272, 2019.

[41] M. M. Solomon, "On the worst-case performance of some heuristics for the vehicle routing and scheduling problem with time window constraints," *Networks*, vol. 16, no. 2, pp. 161–174, 1986.

[42] S. S. Habashi, C. Salama, A. H. Yousef et al., "Adaptive diversifying hyper-heuristic based approach for timetabling problems," in *Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 259–266, IEEE, Vancouver, Canada, November 2018.

[43] S. Abdullah, E. K. Burke, and B. McCollum, "Using a randomised iterative improvement algorithm with composite

neighbourhood structures for the university course timetabling problem," in *Metaheuristics, Operations Research Computer Science Interfaces Series*, vol. 39, pp. 153–169, Springer US, Boston, MA, USA, 2007.

[44] I. P. Gent, M. Ewan, P. Patrick et al., "An empirical study of dynamic variable ordering heuristics for the CSP," in *Principles and Practice of Constraint Programming CP96, Lecture Notes in Computer Science*, E. Freuder, Ed., vol. 1118, pp. 179–193, Springer, Berlin, Germany, 1996.

[45] D. D. Lewis, "Naive (bayes) at forty: the independence assumption in information retrieval," in *Machine Learning: ECML-98*, pp. 4–15, Springer, Berlin, Germany, 1998.

[46] H. Lourenço, O. Martin, and T. Stützle, "Iterated local search," in *Handbook of Metaheuristics, International Series in Operations Research & Management Science*, F. Glover, G. Kochenberger, and F. S. Hillier, Eds., vol. 57, pp. 320–353, Springer, New York, NY, USA, 2003.

[47] T. El-Ghazali, *Metaheuristics: From Design to Implementation*, Wiley Publishing, Hoboken, NJ, USA, 2009.

[48] F. T. Leighton, "A graph coloring algorithm for large scheduling problems," *Journal of Research of the National Bureau of Standards*, vol. 84, no. 6, pp. 489–506, 1979.

[49] P. Augerat, J. M. Belenguer, E. Benavent et al., "Computational results with a branch and cut code for the capacitated vehicle routing problem," 1995.

[50] N. Christofides, "The vehicle routing problem," in *Combined Optimization*, N. Christofides, A. Mingozzi, P. Toth et al., Eds., Wiley, New York, NY, USA, 1979.

[51] B. L. Golden, E. A. Wasil, J. P. Kelly et al., "The impact of metaheuristics on solving the vehicle routing problem: algorithms, problem sets, and computational results," in *Fleet Management and Logistics*, pp. 33–56, Springer, Berlin, Germany, 1998.

[52] E. Uchoa, D. Pecin, A. Pessoa, M. Poggi, T. Vidal, and A. Subramanian, "New benchmark instances for the capacitated vehicle routing problem," *European Journal of Operational Research*, vol. 257, no. 3, pp. 845–858, 2017.

[53] Z. Hanusz, J. Tarasinska, and W. Zielinski, "Shapiro-Wilk test with known mean," *REVSTAT-Statistical Journal*, vol. 14, no. 1, pp. 89–100, 2016.

Hindawi
Complexity
Volume 2021, Article ID 5529447, 11 pages
https://doi.org/10.1155/2021/5529447

Research Article

Web News Data Extraction Technology Based on Text Keywords

Kun Zhang

School of Communication, Xi'an Peihua University, Xi'an City, China

Correspondence should be addressed to Kun Zhang; zhangkun@peihua.edu.cn

Received 18 January 2021; Revised 2 February 2021; Accepted 1 April 2021; Published 16 April 2021

Academic Editor: Abd E.I.-Baset Hassanien

In order to shorten the time for users to query news on the Internet, this paper studies and designs a network news data extraction technology, which can obtain the main news information through the extraction of news text keywords. Firstly, the TF-IDF keyword extraction algorithm, TextRank keyword extraction algorithm, and LDA keyword extraction algorithm are analyzed to understand the keyword extraction process, and the TF-IDF algorithm is optimized by Zipf's law. By introducing the idea of model fusion, five schemes based on waterfall fusion and parallel combination fusion are designed, and the effects of the five schemes are verified by experiments. It is found that the designed extraction technology has a good effect on network news data extraction. News keyword extraction has a great application prospect, which can provide the basis for the research fields of news key phrases, news abstracts, and so on.

1. Introduction

With the development of Internet technology, online news is growing exponentially, and users can get the latest news at home and abroad in real time through mobile phones and other mobile terminals [1]. But in the Internet, the news content is uneven, there are a lot of induced content, lack of news authenticity, it is difficult for users to accurately find the actual needs of the content in the massive network news. At this time, a kind of network news data extraction technology based on text keywords is designed, which can help users accurately locate valuable news [2]. Keywords in text keywords refer to a kind of refined vocabulary to obtain information. The traditional manual keyword extraction work cannot meet the needs of big data news text at this stage, and automatic text keyword extraction technology is imperative [3]. Online news text automatic keyword extraction technology can save users' reading time and, at the same time, assist users to screen out junk news and quickly obtain news content [4]. In view of this, this research will develop a kind of network news text data extraction technology based on text keywords, so as to enhance the quality of news reading and save users' time.

In order to reduce the dependence of Web text keyword extraction on large annotated text corpus, Campos et al. designed an unsupervised automatic keyword extraction method to extract keywords from a single document through multiple local features [5]. Li and other scholars proposed to infer topic distribution bias labels by the topic model algorithm and construct a weighted graph by using random walk algorithm, introduce offset hash random walk on a weighted graph, and extract text keywords by combining labels [6]. Kolokas and other researchers designed a text keyword extraction method based on a recurrent neural network. The network model can map the text keyword sequence to the whole text and perform continuous re-representation of keywords, which has a good keyword extraction effect [7]. Onan and other scholars have investigated the most commonly used keyword extraction methods, such as the measure based keyword extraction method, word frequency-inverse sentence frequency based keyword extraction method, cooccurrence statistical information based keyword extraction method, eccentric keyword extraction method, and TextRank algorithm. After that, they proposed the combination of keyword based text document representation and ensemble learning, which can significantly improve the efficiency of keyword extraction and improve the prediction performance of the text classification scheme to extend the performance [8]. With the continuous improvement of the level of science and technology, the number of on-board

applications such as car navigation has increased, and the amount of social data of cars has increased sharply. Cloud based vehicle data processing methods have emerged. In order to protect the data from being accessed, Yang et al. proposed a keyword extraction measurement method based on specific text spatial distribution, which requests access through the extracted keyword index, so as to achieve data encryption protection [9].

The progress of digital technology has fundamentally affected the society. Hofmann and other researchers integrate text data from different information sources through text mining technology, process it into an analytical and readable relationship network between technologies, and then study the dynamic system of related technologies [10]. Sapozhnikova and other scholars use the convolutional neural network to classify the news information text of Internet information portal and realize the semantic preprocessing of the text through the open word2vec model. In the network news data extraction, the classification accuracy reaches 84% [11]. With the popularity of social media, users' travel preferences can be obtained from users' historical mobile records on social media. Wen et al. designed a representative travel path framework based on keyword perception, extracted knowledge from users' historical mobile records, and successfully completed the travel route recommendation experiment [12]. Ranjan and Prasad used semantic keywords and BPlion neural network algorithm to automatically classify text. Through experiments on data sets, the results show that the accuracy of text classification can reach 90.9% [13]. In the era of big data, with the rapid increase of network digital resources, short text resources show great vitality. Wang and his team analyzed the classification of Chinese short text under low granularity features (keywords) by comparing the classification ability of different Chinese fragments [14].

These methods have made great progress in keyword extraction, text classification, retrieval, and so on, but there is a lack of relevant research on network news data extraction technology. Although many researchers have designed different types of text keyword extraction methods, the method has strong pertinence and cannot be directly used for news data extraction. In order to save time for users to get the news, this paper designs a kind of network news data extraction technology based on text keywords on the basis of considering the characteristics of network news.

2. Design of Network News Data Extraction Technology

2.1. Keyword Extraction Scheme of Network News Text. With the popularity of search engines and social networks, the way people get information has changed, and the Internet has become an important position for information sharing. Major news portals have launched news mobile clients, resulting in a surge of online news data [15].

As can be seen from Figure 1, the number of mobile Internet news users has increased from 366.51 million in 2013 to 660.2 million in 2019; as of June 2019, the number of Internet news users in China has reached 686 million, an

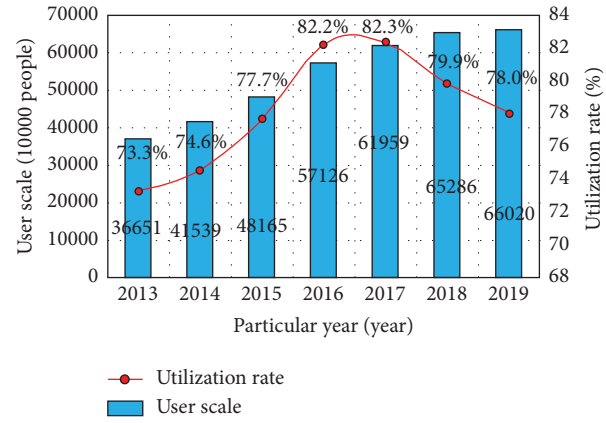


Figure 1: Scale and utilization rate of mobile Internet news users from 2013 to the first half of 2019.

increase of 11.14 million compared with the end of 2018. Through the analysis of user behavior logs, infer user reading preferences, and then push network news to different users so that users have stickiness to the news client. How to accurately extract network news data and achieve "precise and accurate" news push is a powerful tool to save users' time, improve users' reading quality, and improve users' stickiness to news clients [16].

Network news usually focuses on reporting some social events. Generally, only a few keywords are needed to let users understand the main content of the news. Therefore, the extraction of network news data can be summarized as the extraction of keywords in the network news text [17]. Keyword extraction methods are divided into supervised extraction and unsupervised extraction according to whether training samples are needed. This paper mainly studies unsupervised extraction methods, including term frequency–inverse document frequency (TF-IDF) extraction algorithm, TextRank algorithm, and LDA (late Dirichlet allocation) topic model algorithm [18]:

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{N_j}. \tag{1}$$

Formula (1) is the calculation formula of term frequency (TF), where i, j refer to the word and the text corresponding to the word respectively; n_{ij} refers to the number of times the word I appears in text J; the total number of words in text J is represented by N_j:

$$\mathrm{IDF}_i = \log \frac{N}{n_i + 1}. \tag{2}$$

Formula (2) is the calculation formula of inverse document frequency (IDF), the total number of texts is N, and the total number of texts containing word I in the corpus is n_i:

$$\mathrm{TF} - \mathrm{IDF}_i = \mathrm{TF}_{ij} \times \mathrm{IDF}_i. \tag{3}$$

Formula (3) is the calculation formula of TF-IDF. It can be seen that the larger the TF-IDF value of a word i is, the more likely it is to be a keyword of text j. TextRank algorithm

divides the text into several constituent words and constructs the word graph model. Take the automobile network news as an example; see Figure 2 for details.

According to the network diagram shown in Figure 2, the degree of connection between words is explored, the words are scored, and the keywords are obtained by ranking the scores. Set the constructed word graph model as $G = (V, E)$, which is a set of vertices and edges, so the set of all vertices and the set of all edges on the network graph G are represented by V and E in turn. The set of vertices that any vertex v_i points to is $\mathrm{In}(v_i)$, the set that points to other points is $\mathrm{Out}(v_i)$, and $(v_i, v_j) \in E$:

$$H(v_i) = (1-d) + d * \sum_{j \in In(v_i)} \frac{1}{|\mathrm{Out}(v_i)|} H(v_j). \quad (4)$$

Equation (4) is the scoring formula of the vertex v_i of the weighted graph, d is the damping coefficient, and d is 0.85:

$$H(v_i) = (1-d) + d * \sum_{j \in \mathrm{In}(v_i)} \frac{w_{ji}}{\sum_{v_k \in \mathrm{Out}(v_i)} w_{jk}} H(v_j). \quad (5)$$

Equation (5) is the scoring formula when there is a fixed weight between two vertices, where w_{ji} is the weight between v_i and v_j, d is the damping coefficient, and d is 0.85. LDA topic model algorithm combines words and documents which are not directly related by topic to fit the distribution of word text topic.

The original TF-IDF algorithm has the disadvantages of low extraction efficiency and poor extraction accuracy. This paper proposes to introduce Zipf's law and chi-square test to improve the original TF-IDF algorithm, in which Zipf's law is responsible for obtaining weights of different frequencies, and chi-square test is used for keyword extraction. When there are m words in the longer text j, the words with more times appear in the first place and the words with fewer times appear in the second place. The words are ranked by the rank number (word rank) r. When the number of words is n_m, there is $n_m \times r = C$, and C is a constant fluctuating around a fixed value. Most of the online news is in the form of short text, and the frequency of the same word in a single text j is not more than 5 times. The same frequency words are sorted by the maximum method [19, 20]:

$$I_n = r_n - r_{n+1}. \quad (6)$$

Formula (6) is the formula for calculating the number of words I_n with the same frequency, and the value of r_n is the word rank. When the number of words $n_m \le 5$, there is $I_n = (D/n) - (D/(n+1)) = (D/n(n+1)), \quad n \le 5,$ where $D = r_n \times n$, n are word frequency. The proportion of each word frequency in the same text can be counted by I_n/M, and $I_n/M = 1/[n(n+1)]$ and m are approximate constants of the product of word rank and n_m.

It can be seen from Table 1 that, with the increase of n, the value of I_n/M decreases. Because the importance of low-frequency words in short news texts is very low, when extracting network news data, we can first judge whether the word frequency of each word is greater than 1, and if so, calculate the IDF value:

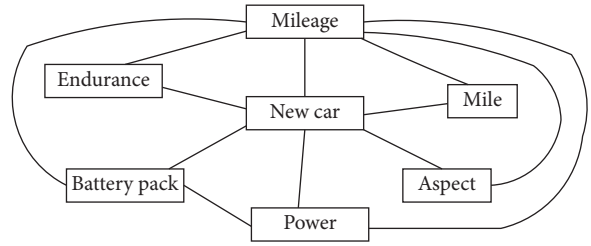


Figure 2: Representation of text graph model.

Table 1: The relation table of I_n/M value with n value.

n	1	2	3	4	5
I_n/M	1/2	1/6	1/12	1/20	1/30

$$\chi^2 = \sum \frac{(A-E)^2}{E}. \quad (7)$$

Equation (7) is the expression of the chi square test χ^2. When χ^2 (degree of deviation) is very small, it is judged as error, where A and E refer to the actual value and the theoretical value, respectively:

$$\mathrm{TF-IDF}-K = \mathrm{TF} \times \mathrm{IDF} \times \log K. \quad (8)$$

Equation (8) is based on the principle of Zipf's law and the chi-square test (tf-idf-k), and there is a chi-square value $K = \chi^2$.

As shown in Figure 3, first preprocess the news text through Jieba word segmentation, then filter the stop words, and count the number of times each word appears, remove the words with word frequency of 1, calculate the TF-IDF value and chi-square value K, multiply to obtain the tf-idf-k value, and arrange them in descending order. The top words are the output results, that is, the text keywords.

2.2. Network News Data Extraction Scheme Based on Model Fusion. Model fusion can significantly improve the accuracy of network news data. Two model fusion schemes are proposed: the first is waterfall fusion and the second is parallel combination fusion.

As shown in Figure 4, waterfall fusion is in the form of cascading multiple algorithm models, using different algorithms for filtering, so as to get the final result. In the process of waterfall fusion, the previous activity is taken as the input, the result filtered by the previous algorithm is taken as the input filtered by the next algorithm, and the candidate results are continuously screened, so as to obtain the final result with less quantity and high quality [21].

As can be seen from Figure 5, parallel combination fusion extracts keywords from the original document through several groups of algorithms and then scores the keywords in the way of parallel voting, so as to select the optimal result. The three extraction algorithms described in Section 2.1 all have defects. TF-IDF relies heavily on corpus and its accuracy is affected by IDF; the TextRank algorithm has too high computational complexity; and the LDA

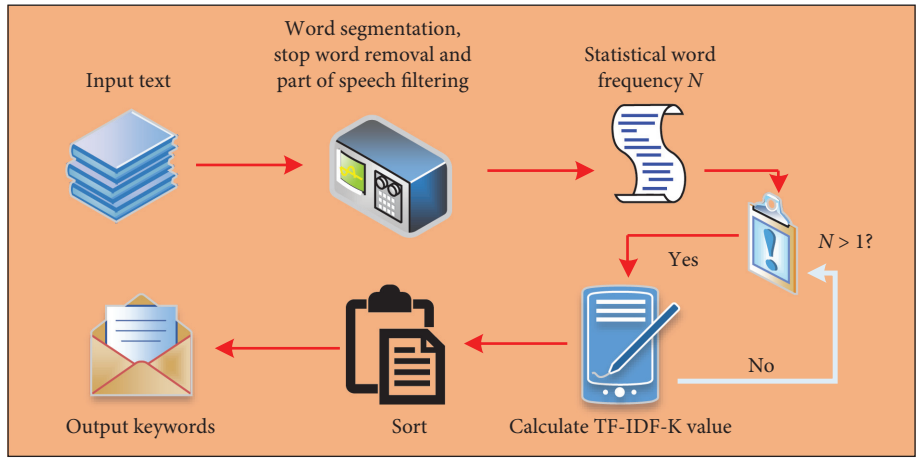


Figure 3: Algorithm flow chart.

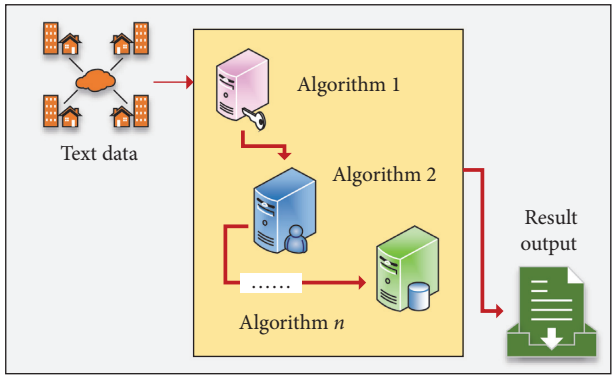


Figure 4: Waterfall fusion flow chart.

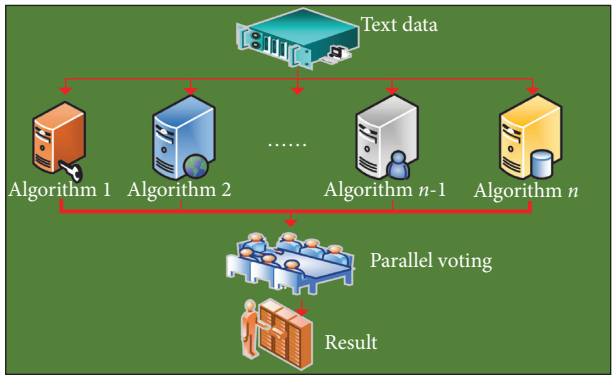


Figure 5: Flow chart of parallel combination fusion.

algorithm cannot feed back accurate document topics. Combined with the characteristics of different fusion models, two kinds of waterfall fusion network news keyword extraction schemes are proposed.

Figure 6(a) is Scheme 1: first, TF-IDF algorithm, and then TextRank algorithm. After word segmentation and stop words removal, the part of speech is filtered and the word frequency n is counted. The TF-IDF value is calculated and sorted according to the TF-IDF value of each word. When the number of times is greater than 50, reorder; when the number of times is not greater than 50, calculate the TextRank value, sort, and output keywords according to the serial number of each word. Figure 6(b) is Scheme 2: first, TF-IDF algorithm, and then LDA topic model algorithm.

The three design schemes in Figure 7 do not consider the sequence. Scheme 1 in Figure 7(a) is the parallel combination of the TF-IDF algorithm and the TextRank algorithm; Scheme 2 in Figure 7(b) is the parallel combination of the TF-IDF algorithm and LDA Algorithm; Scheme 3 in Figure 7(c) is the parallel combination of the LDA algorithm and the TextRank algorithm. The general process of the three schemes can be summarized as follows: input the network news text, process the text by word segmentation, stop words removal, part of speech filtering, and count the word frequency n. The two parallel algorithms sort the words at the same time, list the candidate keywords, output the final keywords by using the parallel fusion method, and complete the key information extraction of network news [22]. In this study, the accuracy rate, recall rate, and F1 value are used to evaluate the effect of online news keyword extraction. The expression of accuracy is shown in

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

$$\begin{aligned} F1 &= \frac{1}{(((1/\text{precision}) + (1/\text{recall}))/2)} \\ &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{11}$$

Formulae (10) and (11) are the expressions of recall rate and F1 value in turn, where TP refers to the situation where the label is a positive sample and the prediction is a positive sample; FN refers to the situation where the label is a positive sample but the prediction is a negative sample; FP refers to the situation where the label is a negative sample but the prediction is a positive sample; TN refers to the situation

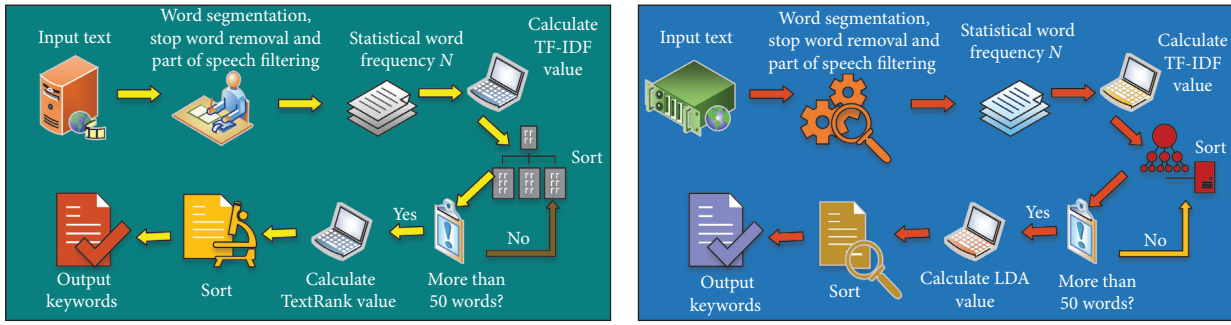


FIGURE 6: Two methods of keyword extraction for online news based on waterfall fusion. (a) Scheme 1 flow chart. (b) Scheme 2 flow chart.

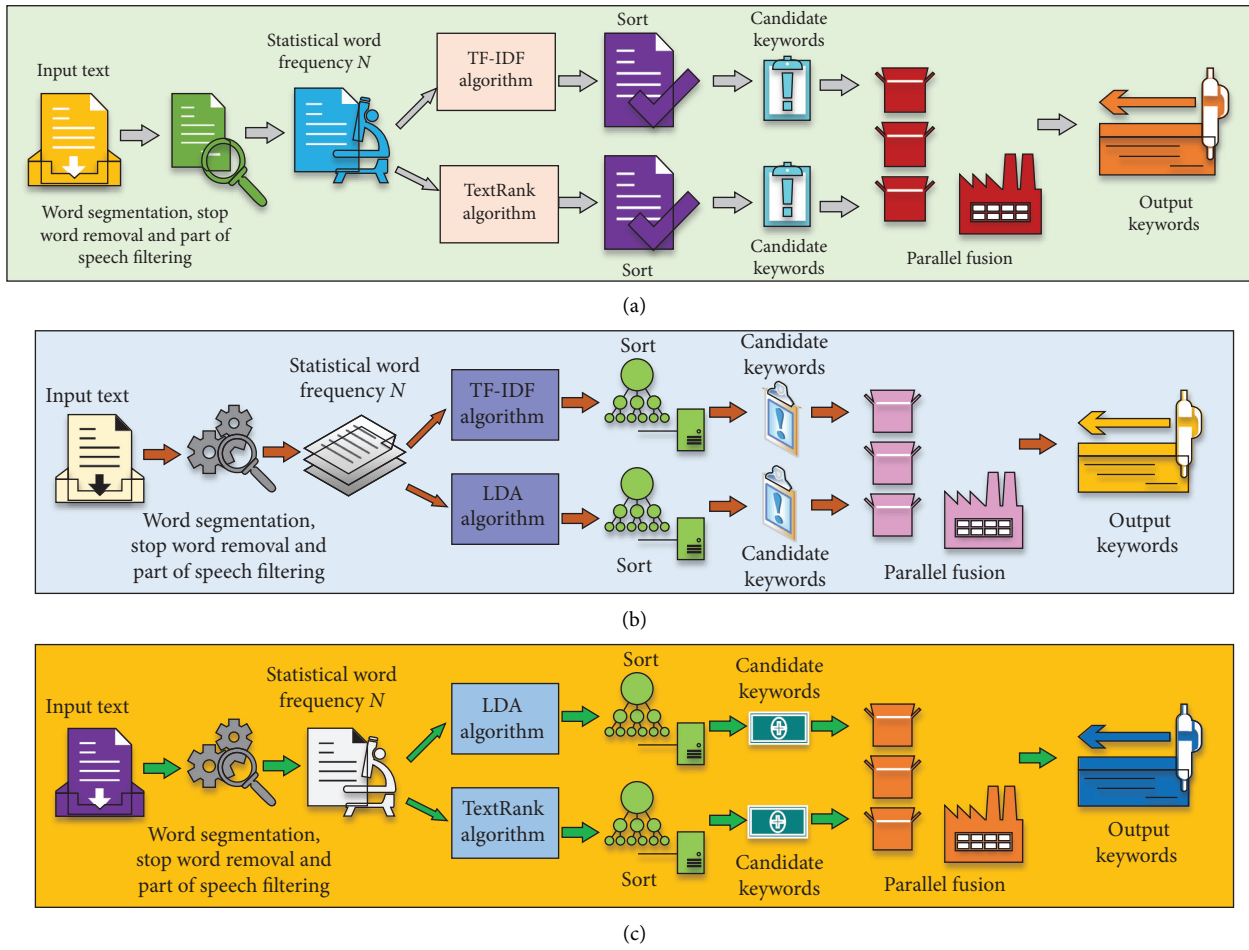


FIGURE 7: Design of three kinds of parallel fusion network news keyword extraction scheme. (a) Scheme 1 flow chart. (b) Scheme 2 flow chart. (c) Scheme 3 flow chart.

where the label is a negative sample and the prediction is a negative sample.

3. Application Effect and Discussion of Network News Data Extraction Technology

3.1. Practical Application Effect of Network News Data Extraction Technology. In order to verify the application effect of the network news data extraction technology proposed in this study, the next stage is the experimental analysis of different schemes. In the experiment, windows 10 system is selected as the experimental operating system, i7 processor is used, and memory is 16g; pychar + python3.6 is selected as the development tool; and 100 network news of ten categories are selected to carry out the experiment.

In the waterfall fusion experiment design, Scheme 1 uses TF-IDF algorithm to extract keywords through the TextRank algorithm; Scheme 2 also uses the TF-IDF algorithm to extract keywords through LDA topic model. The specific experimental results are shown in Table 2.

In Table 2, "Several Private Kindergartens in South Korea Are Involved in Corruption: Embezzling Operating Expenses to Buy Valuable Jewelry" is selected as the extraction object of online news. And through the two schemes for news important information extraction, it can be seen that compared with the standard keywords, it is obvious that the network news keywords extracted by Scheme 1 are closer to the gold standard than those extracted by Scheme 2; that is to say, Scheme 1 has better network news information extraction performance. Select 100 network news texts of ten categories, detect all network news keywords through Scheme 1 and Scheme 2, respectively, and record the time consumed by the two schemes when extracting the key information of the same network news, as shown in Figure 8.

Figure 8 shows the time taken by the two schemes to extract different types of network news keywords. From the analysis of the time taken to extract the key information (keywords) of 10 kinds of network news, the overall time taken to extract keywords in Scheme 1 is much less than that in Scheme 2; the average time taken to extract keywords in Scheme 1 is 43.87 s, and that in Scheme 2 is 138.74 s. The above results show that Scheme 1 has obvious advantages in the time consumption of key information extraction of network news. Ten categories of 100 online news are selected as the experimental objects to compare the recall and accuracy of the two waterfall fusion algorithms. The specific results are shown in Figure 9.

Figure 9(a) shows the recall comparison results of the two waterfall fusion algorithms. On the whole, the recall rate of Scheme 1 is higher than that of Scheme 2; the average recall rate of Scheme 1 is 0.47, and that of Scheme 2 is 0.34, which is 0.13 lower than that of Scheme 1. Figure 9(b) shows the accuracy comparison results of the two waterfall fusion algorithms. On the whole, the accuracy of Scheme 1 is higher than that of Scheme 2; the average accuracies of Scheme 1 and Scheme 2 are 0.38 and 0.31, respectively, and the average accuracy of Scheme 1 is 0.07 higher than that of Scheme 2. The above structure shows that the key information (keywords) acquisition performance of Scheme 1 is better than that of Scheme 2.

In Table 3, online news titled "Japanese College Students' Gathering Led to the Collapse of the Apartment Floor and 30 People Injured" is selected as the experimental object, and three different parallel combination fusion algorithms are adopted to extract the key information of the news. Compared with standard keywords, Scheme 1 (TF-IDF algorithm and TextRank algorithm combined in parallel) is better than Scheme 2 and Scheme 3 in extracting key information of online news. Then, three different parallel combination fusion schemes are used to extract keywords from ten groups of different categories of 100 online news, the specific time consumed by different schemes in extracting key information of different categories of online news is counted, and the average value is calculated. See Figure 10 for details.

Figure 10 shows that the average time consumption of Scheme 2 (parallel combination of LDA topic model algorithm and TF-IDF algorithm) in extracting key information of online news is 92.19 s; Scheme 3 (parallel combination of TextRank algorithm and LDA topic algorithm) in extracting key information of online news is 140.78 s; Scheme 1 (TF-IDF algorithm and TextRank algorithm combined in parallel) extracts the key information of network news, the average time of key information extraction is only 44.77 s. Next, compare the recall rate and accuracy rate of three parallel combination fusion schemes in keyword extraction, and realize the quality comparative analysis of three different parallel combination fusion schemes, as shown in Figure 11.

According to Figure 11(a), the average recall rate of Scheme 2 (parallel combination of LDA topic model algorithm and TF-IDF algorithm) is 0.34; Scheme 3 (parallel combination of TextRank algorithm and LDA topic algorithm) is 0.22; and Scheme 1 (parallel combination of TF-IDF algorithm and TextRank algorithm) is 0.54. From the overall situation of recall rate, the recall rate of Scheme 1 is better than that of Scheme 2 and Scheme 3. Figure 11(b) shows that the average accuracy of Scheme 2 is 0.29, that of Scheme 3 is 0.21, and that of Scheme 1 is 0.35.

In Figure 12, "hot" refers to the popular recommendation method, "$I*$" and "$I*$" refer to the method based on waterfall fusion scheme and the method based on parallel combination fusion scheme, respectively. As can be seen from Figure 12, the effect of popular recommendation is the worst. This is because the popular recommendation method only finds out the popular news list according to the news browsing data of the previous day, filters the news list that users have not browsed, and directly recommends.

3.2. Discussion on Experimental Results. News keyword extraction can help users distinguish junk news and get news content quickly, which has great application prospects. Model fusion is a high-performance classifier composed of multiple classifiers [23]. The research adopts parallel combination fusion technology and designs network news data extraction technology, which can not only realize the accurate extraction of keywords but also greatly shorten the operation time and reduce the budget complexity. For the experimental results of parallel combination fusion extraction, it needs to be carried out from three aspects: intuitive comparative analysis, time comparative analysis, and quality comparative analysis. Among them, intuitive comparative analysis refers to the comparative analysis of the experimental results and the gold standard; time comparative analysis refers to the comparative analysis of the time of extracting network news keywords from the three schemes; quality comparative analysis refers to the respective standard of the three schemes. The accuracy rate and recall rate were compared and analyzed.

From the analysis of the key information extraction time of network news, the average time-consuming of LDA topic model algorithm and TF-IDF algorithm parallel combination is 92.19 s; the average time-consuming of the TF-IDF algorithm and TextRank algorithm parallel combination is

Table 2: Comparison of experimental results.

Title	Standard keywords	Scheme 1	Scheme 2
Corruption Involved in Several Private Kindergartens in South Korea: Embezzling Operating Expenses to Buy Valuable Jewelry	South Korea Kindergarten Corruption related Examination Appropriation	South Korea Examination Appropriation Official Embezzlement	Kindergarten Subsidy Investigate Examination A citizen

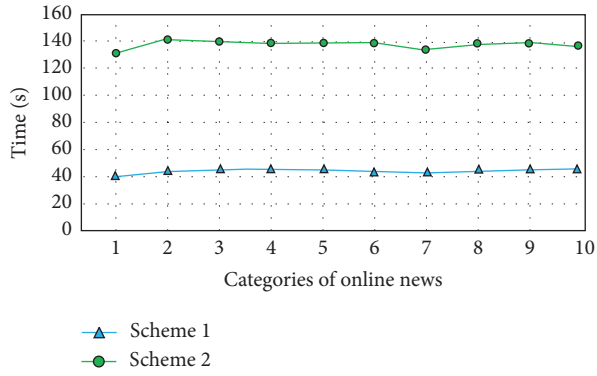


Figure 8: Efficiency comparison chart of keyword extraction.

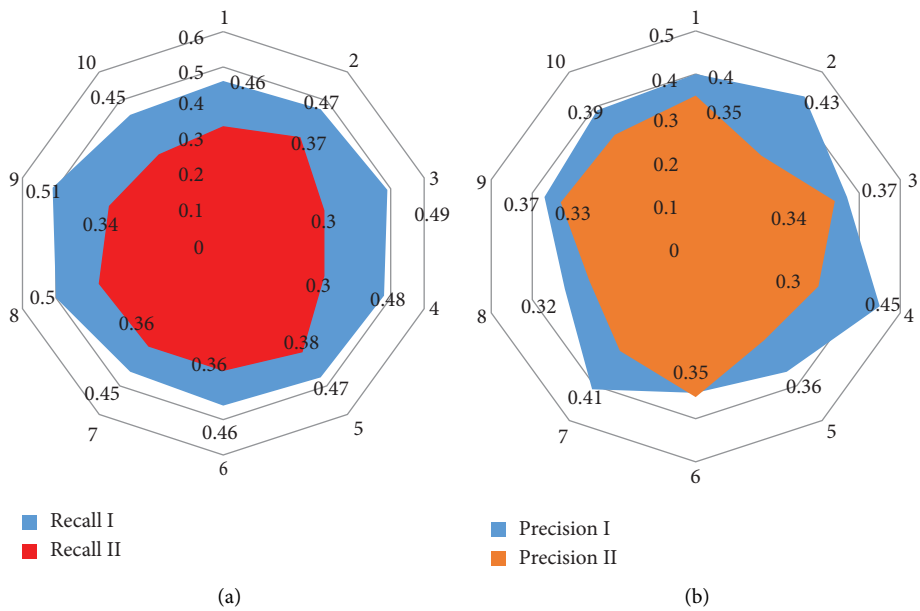


Figure 9: Comparison of recall and accuracy of two waterfall fusion algorithms. (a) Recall comparison. (b) Accuracy comparison.

only 44.77 s. In other words, the TF-IDF algorithm and TextRank algorithm have better time efficiency in network news data extraction. Compared with TF-IDF algorithm, TextRank algorithm, and LDA topic model algorithm, the time consumption of the proposed algorithm on key information of online news is significantly reduced. The average accuracy of the parallel combination of the LDA topic model algorithm and TF-IDF algorithm is 0.29; the average accuracy of the parallel combination of the TextRank algorithm and LDA topic algorithm is 0.21; the average accuracy of the parallel combination of TF-IDF algorithm and TextRank algorithm is 0.35. It can be seen that the accuracy of the TF-IDF algorithm and TextRank algorithm parallel combination is the best, suggesting that researchers can start

Table 3: Comparison of three parallel combination fusion experiments.

Title	Standard keywords	Scheme 1	Scheme 2	Scheme 3
30 People Injured in Apartment Floor Collapse Caused by Japanese College Students' Party	Japan College student University Apartment Accident	University Accident Japan Happen Hold	University Accident Investigation Indoor Student	Accident University Play Happen Party

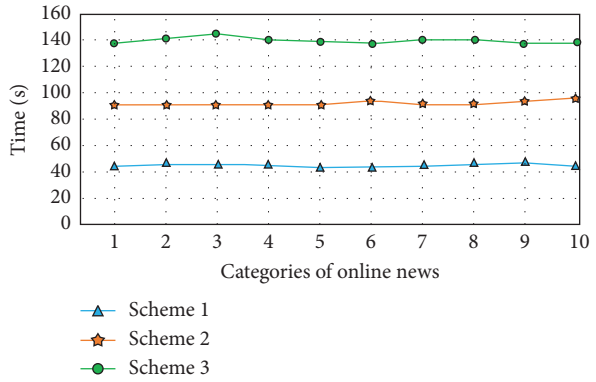
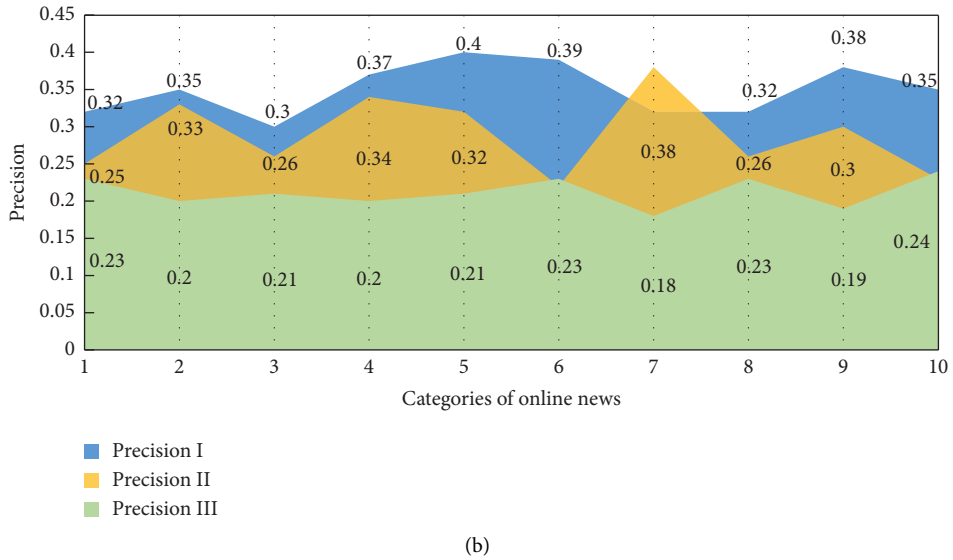


Figure 10: Comparison of three parallel schemes for keyword extraction.

(a)

Figure 11: Continued.



(b)

Figure 11: Quality comparative analysis of three parallel combination fusion schemes. (a) Recall comparison of three parallel combination fusion schemes. (b) Accuracy comparison of three parallel combination fusion schemes.

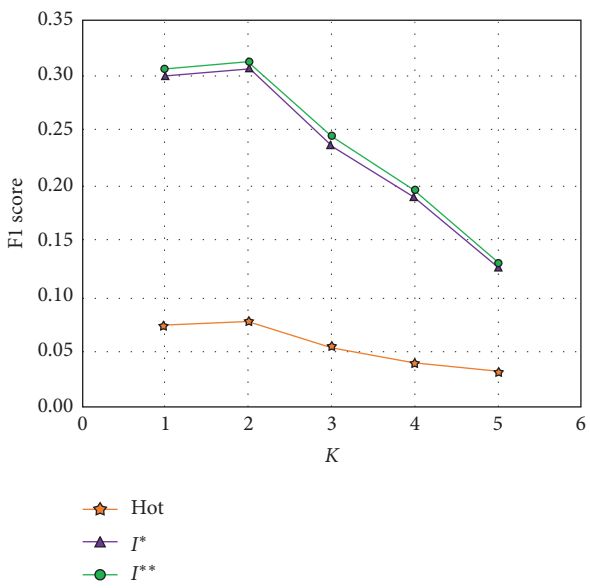


Figure 12: Comparison of news recommendation results under different key information extraction algorithms.

from this combination to further optimize the key information extraction technology of network news [24].

In addition to the consideration of extraction time and accuracy, whether the extracted information meets the needs of users is also very important. In order to compare the application effect of the proposed scheme in the network news data extraction, the research will be based on the above two advantage schemes: first, the TF-IDF algorithm and then TextRank algorithm waterfall fusion scheme, TF-IDF algorithm, and TextRank algorithm parallel combination fusion scheme. The extracted network news keywords are combined with the analysis of different users' historical browsing behavior trajectory, and the corresponding users have analyzed Recommend sex news. And through a recommendation contest platform to score the recommendation effect, compare the news recommendation effect of popular recommendation method, method based on waterfall fusion scheme, and method based on parallel combination fusion scheme and then evaluate the effect of network news data extraction under different methods. From the experimental results, we can see that the popular recommendation method lacks the extraction of network news keywords, and the content lacks pertinence when recommending news to users. Secondly, based on the waterfall fusion scheme, this method can quickly grasp the key information contained in the news through the extraction of network news keywords, improve the pertinence of news recommendation, and effectively shorten the energy consumption and running time in the process of news recommendation. However, the effect of news recommendation based on the waterfall fusion scheme is slightly worse than that based on the parallel combination fusion scheme. This is because in the parallel combination fusion scheme, the two algorithms (TF-IDF algorithm and TextRank algorithm) are not in order, and the corresponding recall rate is high. The more the extracted keywords fit the actual content of news, the more the recommended news is easy to be used as Reading interest.

4. Conclusion

With the rapid development of network news, news content presents an uneven phenomenon, media maliciously exaggerate reports, attract traffic phenomenon is common, and it is difficult for users to quickly obtain the required news content from the massive network news. Network news data extraction technology based on news text keyword extraction has become an effective tool to solve this problem. In view of this, this experiment starts with the

unsupervised keyword extraction method and improves three algorithms based on the analysis of TF-IDF algorithm, TextRank algorithm, and LDA topic model algorithm. The TF-IDF algorithm is improved by yuzipf's law and chi-square test, and five different key information extraction schemes are designed by using waterfall fusion algorithm and parallel combination fusion algorithm combined with the above three unsupervised keyword extraction algorithms. In order to verify the key information extraction effect of different schemes, this paper selects 100 network news of ten categories as the keywords extraction object and verifies the key information extraction effect of network news from three aspects of visual comparative analysis, time comparative analysis, and quality comparative analysis. Finally, through the news recommendation contest, the paper compares the key information extraction effect of network news designed in this study from the side. The results show that the designed extraction technology has a good effect on network news data extraction, and the keyword extraction performance of model fusion is higher than that of traditional extraction methods. Although some achievements have been made in this study, Jieba word segmentation is directly used in the keyword preprocessing step, and the advantages of each algorithm model are not maximized. In the future, a voting mechanism will be introduced to maximize the advantages of each algorithm model, so as to give full play to the advantages of each algorithm model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Vanyushkin and L. Graschenko, "Analysis of text collections for the purposes of keyword extraction task," *Journal of Information and Organizational Sciences*, vol. 44, no. 1, pp. 171–184, 2020.

[2] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using Bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2020.

[3] P. Symeonidis, L. Kirjackaja, and M. Zanker, "Session-aware news recommendations using random walks on time-evolving heterogeneous information networks," *User Modeling and User-Adapted Interaction*, vol. 30, no. 4, pp. 1–29, 2020.

[4] K. Rohit Kumar, G. Anurag, and N. Pratik, "FNDNet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[5] T. M. Fagbola, C. S. Thakur, and O. Olugbara, "News article classification using Kolmogorov complexity distance measure and artificial neural network," *International Journal of Technology*, vol. 10, no. 4, pp. 710–720, 2019.

[6] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[7] L. Li, J. Liu, and Y. Sun, "Unsupervised keyword extraction from microblog posts via hashtags," *Journal of Web Engineering*, vol. 17, no. 1-2, pp. 97–124, 2018.

[8] A. Onan, S. Korukoğlu, S. Lu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[9] A. Kolesnikov, P. Kikin, G. Niko, and E. Komissarova, "Natural language processing systems for data extraction and mapping on the basis of unstructured text blocks," *InterCarto. InterGIS*, vol. 26, no. 1, pp. 375–384, 2020.

[10] Z. Yang, H. Yu, J. Tang, and H. Liu, "Toward keyword extraction in constrained information retrieval in vehicle social network," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4285–4294, 2019.

[11] P. Hofmann, R. Keller, and N. Urbach, "Inter-technology relationship networks: arranging technologies through text mining," *Technological Forecasting and Social Change*, vol. 143, pp. 202–213, 2019.

[12] L. E. Sapozhnikova and O. A. Gordeeva, "Text classification using convolutional neural network," *Information Technology and Nanotechnology*, vol. 2416, pp. 219–226, 2019.

[13] Y. T. Wen, J. Yeo, and W. C. Peng, "Efficient keyword-aware representative travel route recommendation," *IEEE Transactions on Knowledge & Data Engineering*, vol. 29, no. 99, pp. 1639–1652, 2017.

[14] N. M. Ranjan and R. S. Prasad, "Automatic text classification using BPLion-neural network and semantic word processing," *Imaging Science Journal the*, vol. 66, no. 2, pp. 1–15, 2017.

[15] H. Wang and S. Deng, "A paper-text perspective," *The Electronic Library*, vol. 35, no. 4, pp. 689–708, 2017.

[16] H.-M. Kim, "The analysis of characteristics and plan to activate the small wedding reported in Internet news," *Journal of the Korea Entertainment Industry Association*, vol. 13, no. 3, pp. 43–54, 2019.

[17] Z. Wang, K. Hahn, Y. Kim, S. Song, and J.-M. Seo, "A news-topic recommender system based on keywords extraction," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4339–4353, 2018.

[18] S. Venkatachalam, L. P. Subbiah, R. Rajendiran, and N. Venkatachalam, "An ontology-based information extraction and summarization of multiple news articles," *International Journal of Information Technology*, vol. 12, no. 2, pp. 547–557, 2020.

[19] H. Mingpan, M. Haigang, and M. Changyong, "Male gibbon loud morning calls conform to Zipf's law of brevity and Menzerath's law: insights into the origin of human language - ScienceDirect," *Animal Behaviour*, vol. 160, pp. 145–155, 2020.

[20] M. A. Helal and M. Mouhoub, "Topic modelling in bangla language: an LDA approach to optimize topics and news classification," *Computer and Information Science*, vol. 11, no. 4, pp. 77–83, 2018.

[21] N. Wen, B. He, Z. Yuan, and Y. Fan, "A object detection algorithm based on pyramid convolutional neural networks (CNN) and feature map fusion model," *Abstracts of the ICA*, vol. 1, p. 1, 2019.

[22] M. Hammad and K. Wang, "Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network," *Computers & Security*, vol. 81, pp. 107–122, 2019.

[23] X. Chen, L. Ke, Q. Du, J. Li, and X. Ding, "Facial expression recognition using kernel entropy component analysis network and DAGSVM," *Complexity*, vol. 2021, no. 2, pp. 1–12, 2021.
[24] X. Chen, H. Chen, and H. Xu, "Vehicle detection based on multifeature extraction and recognition adopting RBF neural network on ADAS system," *Complexity*, vol. 2020, no. 2, pp. 1–11, 2020.

Hindawi
Complexity
Volume 2021, Article ID 5597168, 9 pages
https://doi.org/10.1155/2021/5597168

Research Article

Multidirection Object Detection in Aerial View of Traffic Target under Complex Scenes

Zeqing Zhang,[1,2] Weiwei Lin,[3,4] and Yuqiang Zheng[1]

[1]*School of Information Science and Engineering, Xiamen University, Xiamen, China*
[2]*West Yunnan University of Applied Sciences, Dali, China*
[3]*School of Electronic and Information Engineering, Fujian Polytechnic Normal University, Fuqing, China*
[4]*Engineering Research Center for ICH Digitalization and Multi-source Information Fusion, Fujian Province University, Fuqing, China*

Correspondence should be addressed to Yuqiang Zheng; zhengyuqiang@stu.xmu.edu.cn

Received 30 January 2021; Revised 21 February 2021; Accepted 31 March 2021; Published 12 April 2021

Academic Editor: Abd E.I.-Baset Hassanien

Focusing on DOTA, the multidirectional object dataset in aerial view of vehicles, CMDTD has been proposed. The reason why it is difficult for applying the general object detection algorithm in multidirectional object detection has been analyzed in this paper. Based on this, the detection principle of CMDTD including its backbone network and multidirectional multi-information detection end module has been studied. In addition, in view of the complexity of the scene faced by aerial view of vehicles, a unique data expansion method is proposed. At last, three datasets have been experimented using the CMDTD algorithm, proving that the cascaded multidirectional object detection algorithm with high effectiveness is superior to other methods.

1. Introduction

With the development of deep learning, rapid progress has been made in remote sensing or aerial image processing and analysis [1–5]. However, these methods cannot deal with multidirectional object detection. Different from the traditional object detection, which the detection frame is generally horizontal or vertical giant, the detection frame given by multidirectional object detection can be rectangular in any direction. Recently, a plurality of popular general object detection algorithms has been ineffective on vehicle object detection datasets with the best result reaching only 52.93%. Compared with the experimental result on the DOTA dataset [6], the rotation of the detection object boundary box, a large number of small vehicle objects in aerial view of images, and insufficient utilization of data information are main causes for the poor performance.

To improve the effect of multidirectional aerial view of the vehicle object detection dataset, the CMDTD (a cascaded multidirectional object detection algorithm) algorithm has been utilized with FasterRCNN as the benchmark method through using the cascade idea of Cascade RCNN for reference. Moreover, vehicle objects are classified in a coarse-to-fine manner for fitting boundary prediction via a multi-information cascade output end. Data augmentation training is performed on the classes with few samples in statistical data in order to address the problem of sample imbalance. By using the method, the effect of multidirectional object detection can be effectively improved. Hence, it has certain research value for implementing multidirectional vehicle object detection in complex scenes.

2. Research Models

2.1. ResNeXt Backbone Network. Compared with land vehicles, although there are few classes of aerial vehicles objects, big differences between the classes can be witnessed. In order to extract superior object features, ResNeXt [7], which

is more powerful than ResNet, is selected in the backbone network of CMDTD with the adoption of the feature fusion method of FPN [8]. The submodule structures of ResNet and ResNeXt are shown in Figure 1. The submodule of ResNet is composed of three convolutional layers. Regarding the input feature map X, X will extract new features, i.e., $F(X)$ upon three convolution operations. And then, residual blocks are output upon the superimposition of new and old features. The process can be expressed as

$$Y = X + F(X). \tag{1}$$

According to equation (1), the feature obtained by $F(X)$ is the difference between Y and X, which is also called residual characteristic, and the submodule is called residual block. Based on ResNet, the residual was expanded using ResNeXt.

As shown in Figure 1, three convolutional layers are divided into 32 groups of convolutional combinations with consistent parameter sizes, and the sum is equal to the size of the three original convolutional layers. The feature map X is superimposed, or $\sum_{i=1}^{32} F_i(X)$, after 32 groups of convolution features. On this basis, new features are superimposed with the old features for acquiring output features, which is shown as

$$Y = X + \sum_{i=1}^{32} F_i(X). \tag{2}$$

Although parameters are not increased in the new combination, the complexity of feature transformation is increased, strengthening the expressive ability of network towards features.

2.2. Multi-Information Cascade Output. The output end of the general detector and the multidirection and multi-information detection output end proposed in this paper are demonstrated in Figure 2. Based on Figure 2(a), the output end of the general detector should undergo a class determination and border prediction after acquiring the region proposal in RPN (regional proposal network). Being consistent with the output end of the general detector in the first half part, the multi-information cascade output end proposed in this paper will regress to the horizontal outer border of the object. However, it only determines whether the object is foreground or background in the second stage of class determination. Moreover, the length and width information (length, width, and aspect ratio) of the object can be calculated upon obtaining the horizontal frame of the object. Next, RoI pooling performs a new object feature extraction based on the acquired horizontal frame. Finally, a fine classification of the object is performed in the second FCN (fully connected network) based on the length and width information of the object and extracted features. At the same time, the third FCN predicts the positions of four vertices of the object according to the extracted features, acquiring the quadrilateral bounding box of the object.

Compared with the detection of land vehicles, sizes of objects in aerial view of vehicles are diverse, causing that it becomes more difficult to regress the positions. Preciser object boundary can be obtained in the second stage in comparison to the object proposal region, contributing to the boundary positioning in the future. Regarding class determination, objects of different classes may have the same texture and color features yet in varied sizes. For example, small vehicles and large vehicles with similar color characteristics can be distinguished as per the size and aspect ratio of the object. Hence, introducing the aspect ratio information of the object during the fine classification of the object might improve the accuracy of classification. During implementation, the length, width, and aspect ratio calculated by the horizontal frame can be obtained at the detection end, which can be introduced to the FCN together with features upon reducing by 1000 times.

During training, the end is composed of four losses, including two position losses and two class losses. The position loss is the prediction losses of horizontal boundary and the vertex, while the class loss is the foreground classification loss and the fine classification loss. Among them, L1 smoothing loss and cross entropy loss are applied in the position loss and the classification loss, respectively.

2.3. Prediction of Vertex Information. The positioning object boundary at the detection end is presented in Figure 3. In this paper, the object box has been regressed for three times at the detection end. It is performed in chronological order, namely, object candidate box, object horizontal bounding box, and rotating rectangular box, which are corresponding to the dashed box in light blue, the dotted box in blue, and the rectangular box in green.

The dashed box in light blue is far away from the real boundary of the object. It is because the region proposal network regresses to the object boundary according to the anchor. If the anchor point position deviates from the object boundary, a poor regression effect will be achieved. In the experiment, the aspect ratio of anchor is [1 : 1, 1 : 2, 2 : 1, 1 : 3, 3 : 1] with the base size of 6 pixels, so as to adapt to dense small objects and long objects. The dashed box in blue is the horizontal bounding box of regression in the second stage. This regression is approaching to the real boundary of the object as the proposed position of the object is the horizontal rectangle that is externally connected to the regression object. The rectangular box in green is obtained by regressing the four vertices of the horizontal bounding box, and the predictive equation is shown as follows:

$$\begin{aligned} x_1 &= x + \Delta x; \\ y_1 &= y + \Delta y, \\ \Delta x &= t_{x1} \times w + x; \\ \Delta y &= t_{y1} \times h + y, \end{aligned} \tag{3}$$

where x, y, w, and h are the center coordinates, length, and width of the horizontal predictive frame, respectively. And the fully connected network can obtain the first vertex of the object through regressing t_{x1} and t_{y1}. During training, the point closest to the upper left corner of the horizontal

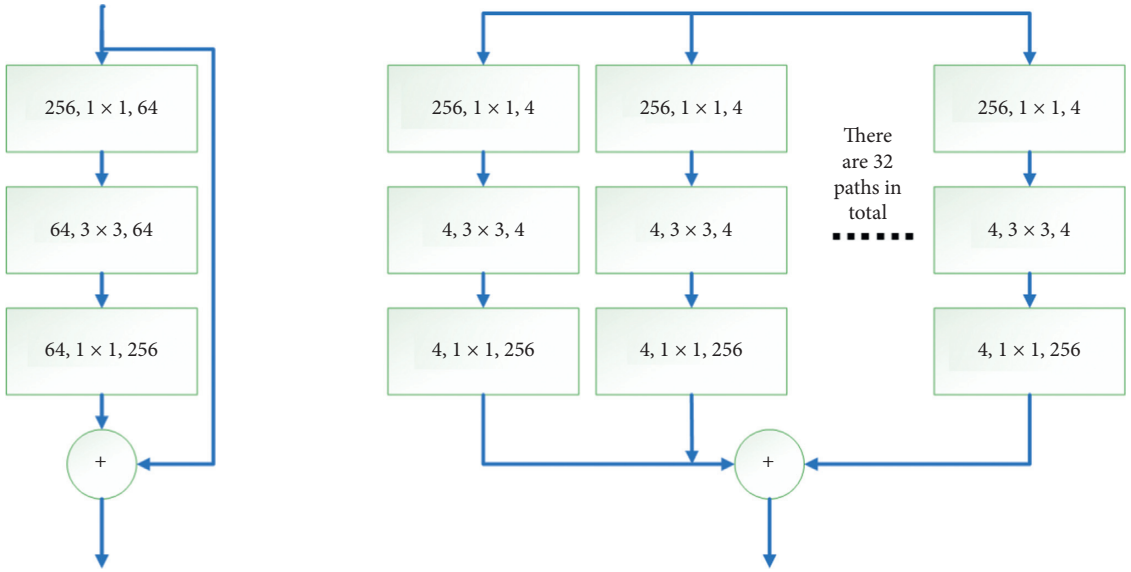


Figure 1: Comparison of residual block structures of ResNet and ResNeXt.

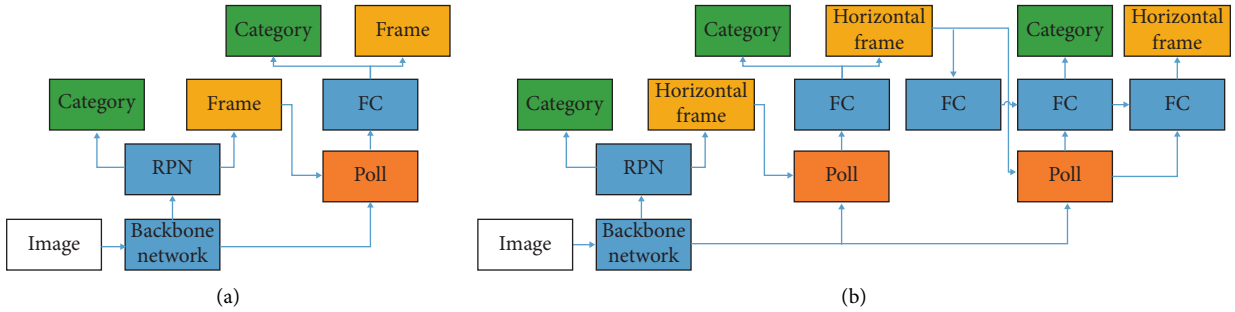


Figure 2: Comparison of general detection end and multidirectional multi-information detection end: (a) the general detection end; (b) the multidirection and multi-information detection end.

bounding box of the object is taken as the first vertex, and the second, third, and fourth vertices are obtained in clockwise order. The real predicted amount can be calculated as

$$t_{x1}^{*} = \frac{(x_1^{*} - x)}{w}, \qquad t_{y1}^{*} = \frac{(y_1^{*} - y)}{h}, \tag{4}$$

where x_1^{*} and y_1^{*} are the real coordinates of the vertex; t_{x1}^{*} and t_{y1}^{*} are the real predictions. And calculation is conducted with the smoothing L1 loss during training. Other coordinate points are calculated in a similar manner.

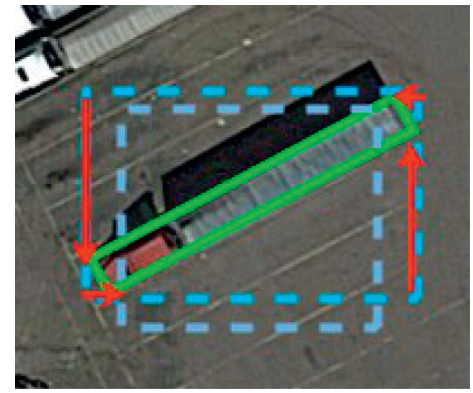


Figure 3: Cascading boundary regression process.

2.4. Unbalanced Data Distribution and Data Augmentation. As can be observed from Figure 4, intensive vehicles and ships can be found due to mass of scenes such as ports and parking lots in aerial view of vehicles images. In the training set after cutting, ships and vehicles have high proportions. In contrast, other classes have extremely proportions.

Images of the 5 classes (including football field, athletic field, rugby field, baseball field, and roundabout) with the least proportions among data are expanded, so as to cope with the problem of unbalanced training data. According to the expanded process presented in Figure 5, horizontal flip, vertical flip, and simultaneously horizontal and vertical flip are performed on images. At this point, data of the class with few proportions have been expanded by 3 times. Since other classes can be witnessed in the augmented image, these classes will be augmented appropriately. And the augmented ratios of all classes are shown in Figure 6.

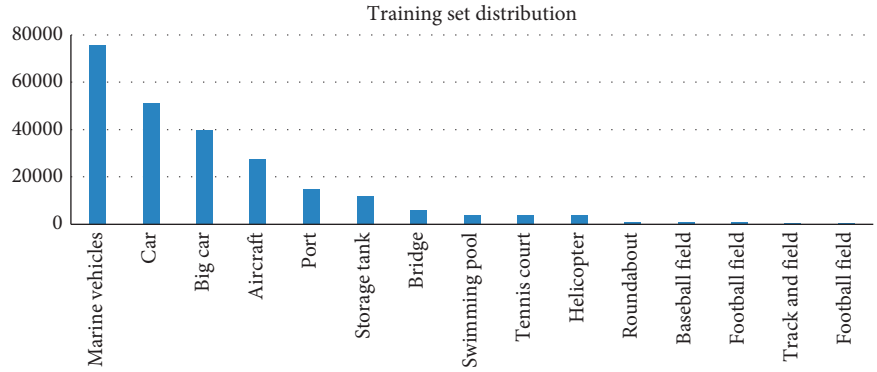


Figure 4: Distribution statistics of the DOTA training dataset.

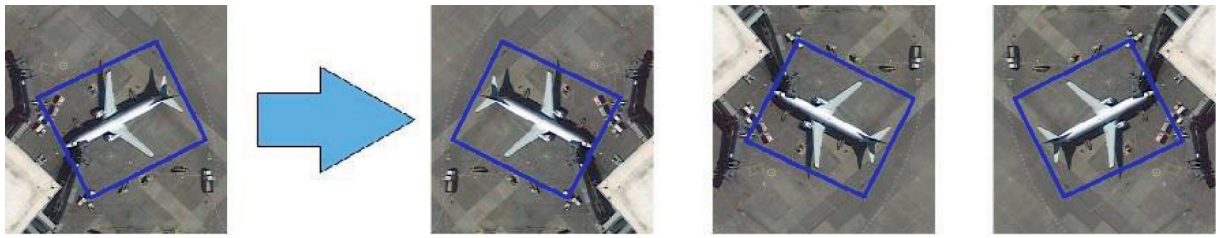


Figure 5: Data augmentation diagram.

3. Experimental Results

3.1. Experimental Dataset and Evaluation Indicators. Verification has been performed on the vehicle-based dataset: DOTA [6] and HRSC2016 [9] in this paper. Moreover, CMDTD also performs verification on the multidirectional scene and ICDAR2015 [10]. As a large-scale aerial view of the vehicle image dataset, the DOTA dataset contains 1,411 training images, 937 test images, and 458 verification images. These image sizes are ranged from 800×800 to 4000×4000, including 15 classes and 188, 282 instance objects. The dataset provides horizontal rectangle labeling and vertex labeling with two detection tasks opened. The first task is a multidirectional object detection task, and the second task is a horizontal object detection task.

HRSC2016 is a dataset of aerial view of maritime vehicle images that are originated from 6 main ports. To be specific, the numbers of training set, verification set, and test set are 436, 181, and 444, respectively, and the image size is ranged from 300×300 to 500×900. The ICDAR2015 dataset is originated from a detection task of the ICDAR 2015 Robust Reading Competition. It is a task that collects images taken in reality. Specially, 1000 out of 1500 images are training images, while the remaining 500 images are test images with the size of 720×1280.

To compare with other methods, CMDTD adopts the calculation standard evaluation method of mAP in DOTA and HRSC2016 and evaluates with F-measure in ICDAR2015. The index is calculated by the recall rate and precision, shown as follows:

$$F - \text{measure} = \frac{1}{(\alpha/\text{Precision}) + ((1-\alpha)/\text{Recall})}, \tag{5}$$

where α is generally set as 0.5.

3.2. Experimental Setup. The experimental environment configured for CMDTD is shown in Table 1.

Concerning the ICDAR2015 dataset, CMDTD is processed through cutting the image into 1088×1088 in a way similar to the DOTA dataset. Besides, the network input size is set to 1088×1088 during training, with training frequency and parameter details consistent with HRSC2016. The input size of the network is set to 720×1280 for testing, so that the image can be input in its original size for testing.

3.3. Result Comparison. The comparison results of five methods such as CMDTD and RRPN [11], ICN [12], and SCRDet [13] in the DOTA dataset task 1 (multidirectional detection) are displayed in Table 2. The mAP indexes of CMDTD are higher than that of other methods, reaching up to 72.81%. In addition, concerning small object detection, the first and the second positions in the detection accuracy of CMDTD belong to ships and cars with 85.5% and 69.51%, respectively.

The comparison results of 5 methods in CMDTD and RFCN [6], ICN [12], and SCRDet [13] in the DOTA dataset task 2 (horizontal detection) are shown in Table 3. The detection result of CMDTD in the horizontal direction is obtained by taking the smallest external rectangle from the

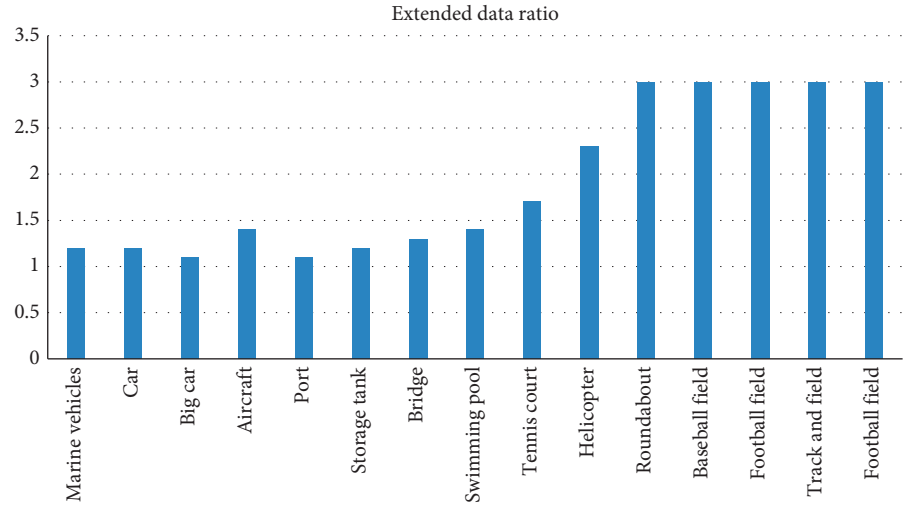


Figure 6: Data augmentation chart. Augmented data ratio: ship, car, big car, airplane, port, storage tank, bridge, swimming pool, tennis court, helicopter, roundabout, baseball field, football field, athletic field, and football field.

Table 1: Configuration of experimental environment.

Operating system	Ubuntu16.04
Kernel version	4.15.0–70-generic
Processor model	Intel(R)core(TM) i7-6700K CPU @ 4.00GHz
Graphics card model	GeForce GTX 1080
Programming language	Python
Frame	Pytorch 1.0.0, mmdetection

Table 2: Comparison results of DOTA task 1 (%).

	RRPN [11]	ICN [12]	Ding et al. [14]	R3Det [15]	SCRDet [13]	CMDTD
Aircraft	88.52	81.40	88.64	89.54	89.98	88.15
Baseball field	71.20	74.30	78.52	81.99	80.65	84.46
Bridge	31.66	47.70	43.44	48.46	52.09	50
Athletic field	59.30	70.30	75.92	62.52	68.36	69.98
Car	51.85	64.90	68.81	70.48	68.36	69.51
Big car	56.19	67.80	73.68	74.29	60.32	73.68
Ship	57.25	70.00	83.59	77.54	72.41	85.5
Tennis court	90.81	90.80	90.74	90.80	90.85	90.35
Rugby field	72.84	79.1	77.27	81.39	87.94	81.09
Storage tank	67.38	78.20	81.46	83.54	86.86	83.52
Football field	56.69	53.60	58.39	61.97	65.02	59.1
Roundabout	52.84	62.90	53.54	59.82	66.68	65.33
Port	53.08	67.00	62.83	65.44	66.25	71.64
Swimming pool	51.94	64.20	58.93	67.46	68.24	66.48
Helicopter	53.58	50.20	47.67	60.05	65.21	54.28
mAP	61.01	68.20	69.56	71.69	72.61	72.81

result of the rotating rectangle. Of which, the method proposed in this paper ranks the second with mAP reaching 73.7%, in surpass of most of the other methods.

The comparison results of CMDTD in 6 methods including the ICDAR2015 dataset, CTPN [18], and RRPN [11] are shown in Table 4. These methods are detection methods based on one stage or two stage. The comprehensive score of the method proposed in this paper ranks the second, reaching 83.88%. An excellent performance can be observed.

The comparison results of CMDTD in 7 methods including the HRSC2016 dataset, RRD [22], and R2CNN [20] are demonstrated in Table 5. The mAP index of the method proposed in this paper ranks the top, reaching 89.68%.

3.4. The Influence of Different Modules on Model Results. Influences of different settings on the model results, including the influence of cascade, the influence of location information

Table 3: Comparison result (%) of DOTA task 2.

	R-FCN [16]	FRH [17]	FPN [8]	ICN [12]	SCRDet [13]	CMDTD
Aircraft	79.33	80.32	88.70	90.00	90.18	87.38
Baseball field	44.26	77.55	75.10	77.70	81.88	83.96
Bridge	36.58	32.86	52.60	53.40	55.30	51.33
Athletic field	53.53	68.13	59.20	73.30	73.29	70.01
Car	39.38	53.66	69.40	73.50	72.09	70.06
Big car	34.15	52.49	78.80	65.00	77.65	74.47
Ship	47.29	50.04	84.5	78.20	78.06	85.99
Tennis court	45.66	90.41	90.60	90.80	90.91	90.36
Rugby field	47.74	75.05	81.30	79.10	82.44	71.49
Storage tank	65.84	59.59	82.60	84.80	86.39	83.60
Football field	37.92	57.00	52.50	57.20	64.53	58.26
Roundabout	44.23	49.81	62.10	62.10	63.45	65.57
Port	47.23	61.69	76.60	73.50	75.77	75.60
Swimming pool	50.64	56.46	66.30	70.20	78.21	73.31
Helicopter	34.90	41.85	60.10	58.10	60.11	54.14
MAP	47.24	60.46	72.00	72.50	75.35	73.70

Table 4: ICDAR2015 comparison results (%).

Methods	Recall rate	Accuracy	Comprehensive scoring
CTPN [18]	51.56	74.22	60.85
RRPN [11]	82.17	73.23	77.44
EAST [19]	78.33	83.27	80.72
R2CNN [20]	79.68	85.62	82.54
FOTS RT [21]	85.95	79.83	82.78
R3Det [15]	83.54	86.43	84.96
CMDTD	80.21	87.92	83.88

Table 5: ICDAR2015 comparison results (%).

Methods	mAP
R2CNN [20]	73.07
RRPN [11]	79.08
RetinaNet-H [10]	82.89
RRD [22]	84.30
RetinaNetR [15]	89.18
RoI-Transformer [14]	86.20
R3Det [15]	89.14
CMDTD	89.68

on classification [23], and the influence of data augmentation are investigated by CMDTD on DOTA task 1. The results are shown in Table 6. Experimental results are presented as follows:

For the influence of data augmentation, the mAP of the model trained by CMDTD on the dataset without data augmentation achieved only 72.48%. Furthermore, the detection effects of football field, rugby field, and baseball field are reduced by 5.8%, 3.33%, and 3.29%, respectively.

Detection results of CMDTD in two aerial views of image datasets can be found in Figures 7 and 8. It can be observed from the figure that there are a large number of objects in the DOTA dataset with significant differences in size and aspect ratios. The object of the HRSC2016 dataset is long rectangle. CMDTD can effectively capture objects in various directions [24], which also presents a satisfactory detection effect on small objects [25].

Table 6: Comparison results of different module settings (%).

	No cascade	No location information	No data augmentation	CMDTD
Aircraft	83.60	88.18	88.73	88.15
Baseball field	65.05	82.99	81.17	84.46
Bridge	37.09	50.13	48.87	50.00
Athletic field	66.29	71.63	71.79	69.98
Car	64.09	70.20	69.03	69.51
Big car	60.49	74.11	74.43	73.68
Ship	76.00	85.08	85.39	85.50
Tennis court	86.43	90.65	90.65	90.35
Rugby field	60.67	77.53	77.76	81.09
Storage tank	67.40	83.74	84.07	83.52
Football field	44.23	53.32	53.30	59.10
Roundabout	51.23	66.01	65.29	65.33
Port	59.55	72.02	71.80	71.64
Swimming pool	54.19	66.36	65.89	66.48
Helicopter	37.73	51.38	58.03	54.28
MAP	60.94	72.22	72.48	72.81

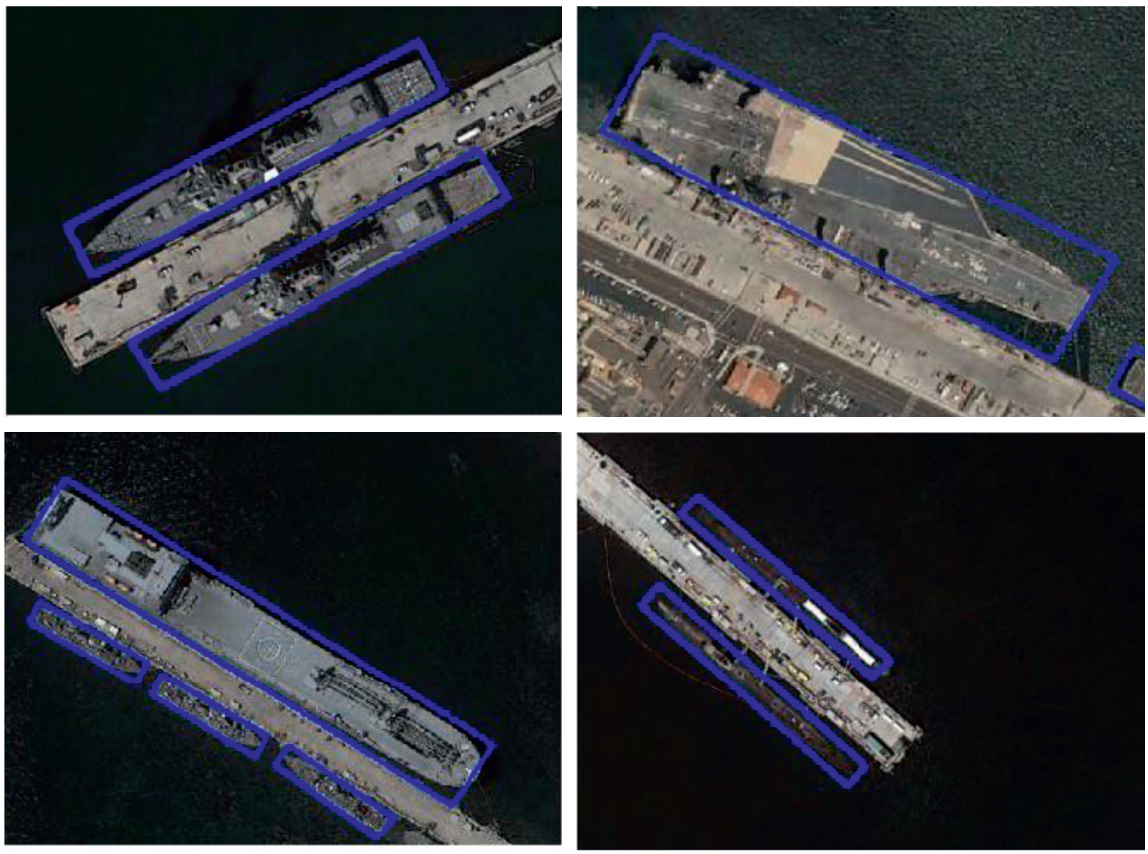


Figure 7: DOTA dataset results.

Figure 8: HRSC2016 dataset results.

4. Conclusion

Firstly, an aerial view based on aerial view of vehicle image object data set is proposed to overcome the shortcomings of the General Object Detection Algorithm in aerial view. On this basis, CMDTD has been proposed. The reason why it is difficult for applying the general object detection algorithm in multidirectional object detection has been analyzed in this paper. Based on this, the detection principle of CMDTD, including its backbone network, multidirectional multi-information detection end module, and aerial view of vehicle data augmentation method has been studied. At last, three datasets have been experimented using the CMDTD algorithm, proving that the cascaded multidirectional object detection algorithm with high effectiveness is superior to other methods.

Data Availability

The data used to support the findings of this study are included within [1, 2].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. Hong, L. Gao, N. Yokoya et al., "More diverse means better: multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 99, pp. 1–15, 2020.

[2] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.

[3] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2019.

[4] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: a novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5146–5158, 2019.

[5] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: an efficient framework for geospatial object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 302–306, 2019.

[6] G. S. Xia, X. Bai, J. Ding et al., "DOTA: a large-scale dataset for object detection in aerial view of images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, Salt Lake City, UT, USA, June 2018.

[7] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, Honolulu, HI, USA, July 2017.

[8] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[9] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some

new baselines," in *Proceedings of the International Conference on Pattern Recognition Appli- Cations and Methods*, pp. 324–331, Porto, Portugal, February 2017.

[10] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou et al., "ICDAR 2015 competition on robust reading," in *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 1156–1160, Tunis, Tunisia, August 2015.

[11] J. Ma, W. Shao, H. Ye et al., "Arbitrary-Oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[12] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proceedings of the Asian Conference on Computer Vision*, pp. 150–165, Perth, Australia, December 2018.

[13] X. Yang, J. Yang,, J. Yan et al., "Scrdet: towards more robust detection for small, cluttered and rotated objects," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8232–8241, Seoul, Korea, November 2019.

[14] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning Roi transformer for detecting oriented objects in aerial view of images," arXiv preprint arXiv:1812.00155, 2018.

[15] X. Yang, Q. Liu, J. Yan, and A. Li, "R3DET: refined single-stage detector with feature refinement for rotating object," arXiv preprint arXiv:1908.05612, 2019.

[16] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional net- works," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 379–387, Barcelona, Spain, December 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 39, no. 6, pp. 91–99, 2015.

[18] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proceedings of the European Conference on Computer Vision*, pp. 56–72, Amsterdam, The Netherlands, October 2016.

[19] X. Zhou, C. Yao, H. Wen et al., "EAST: an efficient and accurate scene text detector," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5551–5560, Honolulu, HI, USA, July 2017.

[20] Y. Jiang, X. Zhu, X. Wang et al., "R2CNN: rotational region CNN for orientation robust scene text detection," arXiv preprint arXiv:1706.09579, 2017.

[21] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: fast oriented text spotting with a unified network," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5676–5685, Salt Lake City, UT, USA, June 2018.

[22] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5909–5918, Salt Lake City, UT, USA, June 2018.

[23] X. Xue, X. Wu, and J. Chen, "Optimizing biomedical ontology alignment through a compact multiobjective particle swarm optimization algorithm driven by knee solution," *Discrete Dynamics in Nature and Society*, vol. 2020, no. 7, pp. 1–10, Article ID 4716286, 2020.

[24] X. Xue and J. Chen, "Using compact evolutionary tabu search algorithm for matching sensor ontologies," *Swarm and Evolutionary Computation*, vol. 48, pp. 25–30, 2019.

[25] Z. Du, J. Pan, S. Chu, H. Luo, and P. Hu, "Quasi-affine transformation evolutionary algorithm with communication schemes for application of RSSI in wireless sensor networks," *IEEE Access*, vol. 8, pp. 8583–8594, 2019.

Hindawi
Complexity
Volume 2021, Article ID 5520223, 11 pages
https://doi.org/10.1155/2021/5520223

Research Article

An Intelligent Passenger Flow Prediction Method for Pricing Strategy and Hotel Operations

Tianyang Wang

City University of Macau, Macau, China

Correspondence should be addressed to Tianyang Wang; t20091100208@cityu.mo

Received 1 February 2021; Revised 11 February 2021; Accepted 3 March 2021; Published 18 March 2021

Academic Editor: Abd E.I.-Baset Hassanien

Hospitality industry plays a crucial role in the development of tourism. Predicting the future demand of a hotel is a key step in the process of hotel revenue management. Hotel passenger flow prediction plays an important role in guiding the formulation of hotel pricing and operating strategies. On the one hand, hotel passenger flow prediction can provide decision support for hotel managers and effectively avoid the waste of hotel resources and loss of revenue caused by the loss of customers. On the other hand, it is the guarantee of the priority occupation of business opportunities by hotel enterprises, which can help hotel enterprises adjust their operation strategies reasonably to better adapt to the market situation. In addition, hotel passenger flow prediction is helpful to judge the overall operating condition of the hotel industry and assess the risk level of the hotel project to be built. Hotel passenger flow is affected by many factors, such as weather, environment, season, holidays, economy, and emergencies, and has the characteristics of complex nonlinear fluctuation. The existing demand predicting methods include linear methods and nonlinear methods. The linear prediction methods rely on the stability of environment and time series, so they cannot completely simulate the complex nonlinear fluctuations characteristics of hotel passenger flow. Traditional nonlinear prediction methods need to improve the prediction accuracy, and they are difficult to deal with the increasing data of hotel passenger flow. Based on the above analysis, this paper constructs a deep learning prediction model based on Long Short-Term Memory (LSTM) to predict the number of actual monthly arrival bookings. The number of actual monthly arrival bookings can reflect the actual monthly passenger flow of a hotel. The prediction model can effectively reduce the loss caused by cancellation or nonarrival of bookings due to various reasons and improve the hotel revenue. The experimental part of this paper is based on the booking demand dataset of a resort hotel in Portugal from July 1, 2015, to August 31, 2017. Artificial neural network (ANN) and support vector regression (SVR) are built as benchmark models to predict the number of actual monthly arrival bookings of this hotel. The experimental results show that, compared with the benchmark models, the LSTM model can effectively improve the prediction ability and provide necessary reference for the hotel's future pricing decision and operation mode arrangement.

1. Introduction

The development of tourism plays an increasingly important role in promoting world economic growth and cultural prosperity. China's tourism has grown steadily in recent years. According to the "Basic Situation of Tourism Market in 2019" released by Chinese Tour Research Institute (http://www.ctaweb.org), the total tourism revenue in China in 2019 was 6.63 trillion yuan, accounting for 11.05% of the total GDP, and the domestic tourism reception scale was 6.006 billion person-times. With the vigorous development of tourism, the demand of tourists for accommodation and environmental requirements are also increasing. Hotel operation is facing both opportunities and challenges.

At present, revenue management is a modern scientific operation management mode widely used in hotel management. It refers to the market-oriented analysis and prediction of the consumer behavior of each market segment in the process of hotel management to determine the optimal price and the optimal stock allocation model, so as to optimize the hotel profitability [1]. Revenue management mainly includes hotel demand prediction, pricing strategy, overbooking strategy, and room allocation strategy [1, 2], which is an important tool for making hotel strategic

decisions. Hotel demand prediction is a key step in the process of revenue management. By collecting historical data and establishing prediction model, hotel managers can understand and master the market demand, so as to make differential pricing and adopt different revenue management strategies in different periods to maximize revenue. Hotel demand prediction can provide basic information for the subsequent development planning and decision-making of the hotel. Therefore, accurate prediction of hotel demand is of great significance to the operation and management of hotels.

Hotel passenger flow not only can reflect the current development trend of a hotel but also is an important embodiment of whether the local tourism is developed. Correct analysis and prediction of hotel passenger flow are the core of hotel demand prediction, which plays a crucial role in hotel operation decisions [3] and is the key to hotel revenue management [4]. Hotel passenger flow is affected by many factors, such as weather, environment, season, holidays, economy, and emergencies [5], and has the characteristics of complex nonlinear fluctuation [6]. Therefore, it is often difficult for hotel managers to accurately estimate the passenger flow, leading to unreasonable decision-making, which leads to the waste of hotel resources and loss of revenue. The randomness of passenger flow makes it difficult to analyze and predict. Finding a suitable prediction method is the key and difficult point of hotel passenger flow prediction. Traditional linear predicting methods cannot accurately fit hotel passenger flow, such as Autoregressive Integrated Moving Average (ARIMA), which is the most widely used time predicting method. Zhang [1] used the ARIMA method to model and predict the hotel occupancy rate data in three regions in the research of predicting hotel occupancy rate. On this basis, she used the Ensemble Empirical Mode Decomposition (EEMD) to make up for the shortcomings of time-series models that cannot accurately capture the characteristics of data fluctuations to improve the prediction accuracy of the ARIMA model. Most of the existing nonlinear prediction methods use traditional machine learning methods to build passenger flow prediction models, which cannot automatically extract feature information and adapt to the growing experimental data. Besides, they are easy to fall into the problem of local optimization and overfitting. Ji [7] used Singular Spectrum Analysis (SSA) to model and predict the hotel monthly occupancy times and income, based on the hotel occupancy history data of Victoria Hotel from January 1980 to June 1995. Wu [3] used the SVR model to predict the monthly occupancy, also based on the statistics of the occupancy situation of Victoria Hotel, so as to provide a meaningful reference for the arrangement of the future business model of Victoria Hotel.

Based on the above analysis, this paper constructs LSTM model with good prediction ability for complex time series to predict the monthly passenger flow of a resort hotel. We predict the number of actual monthly arrival bookings to provide reference for the formulation of hotel pricing strategy and the adjustment of operation mode.

Compared with the existing work, the contribution of this paper can be summarized as follows:

(i) A deep learning algorithm based on automatic extraction of nonlinear complex sequence characteristics is proposed for the first time to predict the number of actual arrival bookings of a hotel.

(ii) A hotel passenger flow prediction deep learning model based on LSTM is constructed to predict the actual arrival bookings in a certain month in the future. This model has better performance than traditional hotel demand prediction methods.

(iii) ANN and SVR are constructed as benchmark models and compared with LSTM model in this paper. Mean absolute error, root mean squared error, and mean absolute percentage error are used as evaluation metrics to evaluate the prediction effect of these three models.

Subsequent parts are organized as follows: Section 2 provides an overview of existing demand prediction methods and previews previous efforts to predict hotel passenger flow. It discusses various types of prediction models. Section 3 analyzes the hotel demand data from a resort hotel in Portugal. Section 4 introduces the LSTM network and the hotel passenger flow prediction model based on LSTM in detail. Section 5 describes the experiments carried out in this paper and analyzes the results; lastly, a summary is discussed in Section 6.

2. Literature Review

As early as the 1960s, experts and scholars had already begun researches on demand prediction methods [8]. The existing demand prediction methods include linear and nonlinear methods [6]. These methods can be divided into four categories: classical time-series methods, econometric methods, machine learning methods, and comprehensive prediction methods [8, 9].

Among the demand prediction methods, classical time-series and econometric methods are representative linear time-series prediction methods. Classical time-series prediction methods are widely used. They mostly use the linear relationship between variables for modeling and predicting [8], which contain a lot of types, mainly including Moving Average, Exponential Smoothing, Differential Autoregression, Regression Prediction Model, and Incremental methods [8, 9]. Andrew et al. [10] built two time-series models based on Box-Jenkins and Exponential Smoothing to predict monthly occupancy rates for hotels and verified that these two models can be very useful in actual hotel operations and other applications such as yield management; Pfeifer et al. [11] described the application of Space-Time ARMA modeling to demand-related data from eight hotels from a single hotel chain in a large US city; Weatherford et al. [4] used data from Choice Hotels and Marriott Hotels to test a variety of prediction methods and to determine the most accurate one; in-depth study using the Marriott Hotel data showed that models based on Exponential Smoothing, pickup, and Moving Average were the most robust; Pan et al. [12] first proposed the value of search query volume data in predicting hotel room demand in the field of tourism and

hospitality research; they used search volume data on five related queries to predict demand for hotel rooms in a specific tourist city and proved that three ARMAX models consistently outperformed their ARMA counterparts; Zhang [1] proposed an ARIMA-EEMD hybrid model and concluded that this model has a better effect than the single ARIMA model in predicting hotel occupancy; Yu [5] constructed a Hotel Management Decision Support System based on Mann-Kendall, Moving Average, Exponential Smoothing, and fuzzy comprehensive evaluation modeling method, which realized the prediction of short-term passenger flow; Yang [13] designed a tourist flow prediction system of resort hotel based on Logistic model and proved that the system has high prediction accuracy, which helps to improve the management and scheduling decision-making ability of resort hotel; Sun [14] realized the demand prediction for cruise revenue management based on non-incremental, classical incremental combined, and advanced incremental combined prediction method. Econometric method can explain the causal relationship between target variables and influencing factors from the perspective of economics, but the influencing factors of econometrics can only be speculated based on theory, which is difficult to clarify [8]. Goh et al. [15] used the time-series Seasonal ARIMA (SARIMA) and Multivariate ARIMA (MARIMA) models to predict a tourism demand in Hong Kong and got highest accuracy compared with eight other time-series models; and then Goh and Law [16] used error correction model to analyze the influence of climate on tourism demand; Choi [17] identified key economic indicators of the hospitality industry in USA and built synthetic indicators to predict the US hotel demands successfully.

Machine learning is the most widely used method in traditional demand prediction. Ji [7] used SSA model to predict the monthly occupancy times and monthly revenue of the hotel based on the hotel occupancy history data of Victoria Hotel from January 1980 to June 1995; Wu [3] used SVR model to predict the monthly occupancy time, also based on the statistical data of Victoria hotel occupancy; Hong et al. [18] presented an SVR model with Chaotic Genetic Algorithm (CGA), namely, SVRCGA, to predict the tourism demands; Rashad et al. [19] developed a fuzzy-rule-based system model for hotel occupancy prediction by analyzing 40 months' time-series data and applying fuzzy c-means clustering algorithm; Sun et al. [20] proposed a new framework integrating machine learning and Internet search index to predict tourist volume and proved that the proposed Kernel Extreme Learning Machine model is more stable and effective according to accuracy and robustness analysis than Least Squares SVR (LSSVR), SVR, Artificial Neural Network (ANN), and ARIMAX.

Comprehensive prediction methods are produced to solve the shortcomings of traditional single prediction methods, including combined prediction, integrated prediction, and hybrid prediction methods [8]. Ke-wei et al. [21] established a combined model based on BP neural network and ARIMA to comprehensively analyze and predict the change trend of China's inbound tourists; Gong and Huang et al. [22] established a demand prediction model based on Grey Theory and Exponential Smoothing method to predict the demand of a certain model of automobile company; according to the nonlinear characteristics of the hotel occupancy rate, ZHANG et al. [23] took Beijing travel-related consumer search data from January 2011 to April 2017 as the input set and constructed a hybrid model integrating consumer search data and SVR. They used Bat Algorithm (BA) to optimize the parameters of SVR, which effectively improves the prediction accuracy of the model.

Through the analysis of the above researches, we find that preview researches were mainly based on linear models to predict hotel demand [1, 4, 5, 11–13, 17, 19]. However, the linear model relies on the stability of the time series and the economic environment, and it is difficult to effectively simulate the nonlinear characteristics of hotel demand. Among the existing nonlinear hotel demand prediction models, SVR [3, 18, 23] is widely used. It is good at using small sample data for prediction and has good processing ability for nonlinear data, but the selection of parameters has a great influence on the prediction results. In addition, the traditional nonlinear prediction models are often unable to deal with the increasing data in practical applications and lack the ability to automatically extract data features.

At present, some scholars also build deep learning models for demand prediction. For example, Chang and Tsai [24] addressed the problem faced by neural network and SVR and proposed the deep learning neural network to predict the tourist arrivals; the result showed that the deep learning applied neural network with feature selection attained the best testing accuracy. Although the deep learning network proposed in this research can carry out feature selection, its prediction accuracy needs to be improved. Zhang et al. [6] did experiments based on a deep learning framework and search index from August 2008 to May 2019 to predict the overnight passenger flows for hotels accommodation in Hainan Province, China, and then constructed an LSTM model incorporating Internet search index to handle the prediction problem in the hotel accommodation demands, which shows good performance in improving the prediction performance. However, not all customers will know the relevant information of the tourist destination through the search engine before booking the hotel, and this method can only predict the number of bookings but cannot predict the actual occupancy, which has certain incompleteness and uncertainty.

The above research methods on hotel demand prediction mainly have one or more of the following problems:

(1) They rely too much on the stability of environment and time series

(2) Characteristics of complex nonlinear fluctuations of hotel passenger flow cannot be extracted automatically

(3) They cannot deal with a large amount of sample data in practical application

(4) It is incomplete and uncertain to estimate hotel passenger flow based on customer search data

Neural network models with multiple hidden layers show strong superiority in learning characteristic information and correlation of complex datasets. Compared with other models, the prediction model based on deep learning methods can make demand prediction more accurately. In addition, LSTM model has more significant advantages than other deep learning models in terms of prediction with sequences as inputs [25, 26].

Based on the above analysis, we use the historical booking demand dataset of a resort hotel in Portugal [27] and establish a deep learning model of hotel passenger flow prediction based on LSTM to predict the number of actual monthly arrival bookings of this hotel. This model can automatically and effectively extract the complex nonlinear characteristics of the hotel passenger flow data and make more accurate analysis and prediction.

3. Booking Demand Data Analysis

The dataset used in this paper is from paper [27]. It contains booking demand data for a resort hotel from July 1, 2015, to August 31, 2017. A total of 31 variables describe the 40060 observations of the resort, including "is_canceled," "lead_time," "arrival_date_year," "arrival_date_month," "arrival_date_week_number," "reservation_status," and "reservation_status_date." "reservation_status" includes three categories: "Canceled," which represents that the booking was canceled by the customer, "Check-Out," which represents that the customer has checked in but already departed, and "No-Show," which represents that the customer did not check in and did inform the hotel of the reason why. The paper predicts the number of actual monthly arrival bookings of hotel, that is, the number of "Check-Out" occurrences of the "reservation_status" in a certain month.

Not all variables are highly associated with the change of "reservation_status." We use CorrelationAttributeEvala as Attribute Evaluator and Ranker as Search Method and select most correlated variables with "reservation_status." We can see them in Table 1.

In Table 1, "deposit_type" indicates if the customer made a deposit to guarantee the booking. This variable has the greatest impact on the final booking status; it can be divided into three categories: "No Deposit," which means that no deposit was made, "Non Refund," which means that a deposit was made in the value of the total stay cost, and "Refundable," which means that a deposit was made with a value under the total cost of stay. "lead_time" represents the number of days that elapsed between the entering date of the booking and the arrival date; "total_of_special_requests" represents the number of special requests made by the customer; "required_car_parking_spaces" represents the number of car parking spaces required; "country" indicates which country the customer comes from; "distribution_channel" represents booking distribution channel, where "TA" represents Travel Agents and "TO" represents Tour Operators; "assigned_room_type" is the room type assigned to the customer; the value of this variable may be different from the room type that the customer has booked, and this is because sometimes the room type assigned will be changed due to overbooking or customer requirements; "booking_changes" represents the number of changes made to booking from the moment the booking was entered until the moment of check-in or cancellation; "customer_type" represents the type of the customer; it has four categories: "Contract," "Group," "Transient," and "Transient-party"; "previous_cancellations" represents the number of cancellations before this booking; it has the least influence on the final booking status of customers.

Table 1: Correlation with "reservation_status."

Variable	Correlation
deposit_type	0.47636
lead_time	0.29455
total_of_special_requests	0.23246
required_car_parking_spaces	0.19226
Country	0.17389
distribution_channel	0.1676
assigned_room_type	0.15666
booking_changes	0.14338
customer_type	0.1254
previous_cancellations	0.11008

Table 2 shows some booking demand data of the resort hotel in July 2017. "d_t," "l_t," "t_o_s_r," "r_c_p_s," "c," "d_c," "a_r_t," "b_c," "c_t," and "p_c" represent 10 variables in turn in Table 1; "a_d_y" represents "arrival_date_year," "a_d_m" represents "arrival_date_month," and "r_s" represents "reservation_status."

From Table 2, we can see that, in July 2017, the hotel had a total of 9 bookings, among which 5 bookings were canceled and 1 booking did not arrive. The number of actual arrival bookings was 3. Usually, the number of hotel bookings can reflect its development trend. However, the booking may be canceled by the customer, or the customer fails to check in. As shown in Table 2, due to the change of customer status, the hotel may have 6 spare rooms, which has a negative impact on the hotel room allocation and pricing and ultimately leads to the loss of hotel revenue. If the number of bookings that actually arrive at the hotel can be estimated in advance, it can help the hotel to prejudge the allocation of rooms, so as to make a more reasonable pricing decision and operation strategy. Therefore, predicting the number of actual monthly arrival bookings has practical application value. The focus of this paper is to establish a prediction model to predict the number of actual arrival bookings in a certain month in the future.

4. Prediction Model

4.1. LSTM Network. What we want to predict in this paper is the number of actual arrival bookings in a certain month in the future, that is, the actual monthly passenger flow; this is a typical nonlinear time-series predicting problem. Time-series prediction analysis refers to using the time characteristics of an event in the past to predict the characteristics of the event in a certain period of time in the future, that is, predicting the future changes of an object according to the existing time-series data.

Table 2: Booking demand data.

a_d_y	a_d_m	d_t	l_t	t_o_s_r	r_c_p_s	c	d_c	a_r_t	b_c	c_t	p_c	r_s
2017	July	No deposit	59	0	0	USA	TA/TO	G	0	Transient	0	Cancelled
2017	July	No deposit	52	0	0	PRT	Corporate	D	0	Transient	0	Cancelled
2017	July	No deposit	17	0	0	PRT	Corporate	A	0	Transient	0	No-show
2017	July	No deposit	52	2	0	GBR	TA/TO	E	0	Transient	0	Check-out
2017	July	No deposit	3	1	0	ESP	TA/TO	A	0	Transient-party	0	Check-out
2017	July	Non-refund	17	0	0	PRT	Corporate	A	0	Transient-party	0	Cancelled
2017	July	Non-refund	0	0	0	GBR	TA/TO	E	0	Transient-party	0	Check-out
2017	July	Refundable	36	1	0	PRT	TA/TO	A	0	Transient	0	Cancelled
2017	July	No deposit	0	1	0	PRT	Direct	D	0	Transient	0	Cancelled

Recurrent Neural Network (RNN) is a commonly used algorithm in time-series prediction. It is a type of neural network with short-term memory capabilities. The connection of RNN can have loop structure, which can improve the accuracy of time behavior modeling in time series, text, audio, and other fields. A connection method introduced by RNN can take the input of hidden layer neurons as output and connect with neurons in the same hidden layer, so that the input can be obtained from the previous time step as part of the incoming neuron information. Therefore, the output of the network is not only related to the current input but also related to the output of the previous moment, which enables RNN to have the short-term memory ability when processing the time-series data of any length. Figure 1 shows the loop structure of RNN.

As can be seen from Figure 1, connections exist not only between neurons in adjacent layers (such as Hidden Layer 1 and Hidden Layer 2) but also between neurons in the same hidden layer in temporal dimension (such as neurons in $t = 0$ time step and neurons in $t = 1$ time step of Hidden layer 1). Suppose that the time step is t, the input of RNN is X_t, the neuron activity value of the hidden layer is Y_t, and the net input vector of the hidden layer is Z_t. Y_t is not only related to X_t but also related to the activity value Y_{t-1} of hidden neurons in the previous time step:

$$\begin{aligned} Z_t &= \tan h\left(W_Z X_t + R_Z Y_{t-1} + b_Z\right), \\ Y_t &= f\left(Z_t\right), \end{aligned} \tag{1}$$

where W is the rectangular input weight matrix, R is the square cyclic weight matrix, b is the bias vector, and f is the nonlinear activation function, usually set to tan$h$ or sigmoid function. In equation (1), when $t = 0$, $Y_t = 0$.

Although RNN has some advantages in the field of time series, its long-term memory ability is weak. The gradient vanishing or gradient explosion that occurs when optimizing RNN in certain time steps makes it difficult to model the long-term structural dependence of the input dataset [6].

LSTM network is the most commonly used variant of RNN, which was proposed by Hochreiter and Schmidhuber in 1997 [25]. LSTM network is better at capturing long-term dependencies than regular RNN models. The LSTM network provides a solution for fusing memory cells, allowing learning of previously forgotten hidden cells, and updating the hidden cells based on new information [6]. LSTM network is composed of many LSTM cells, the main body of

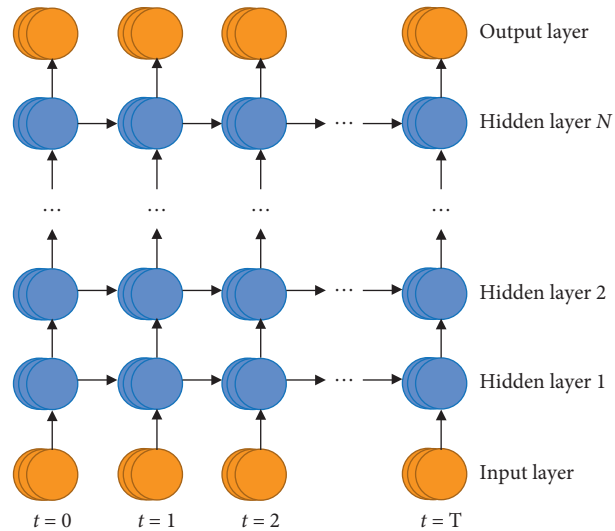


Figure 1: RNN structure unfolded along time axis.

which is LSTM block, including input gate, forgetting gate, output gate, fast input, memory cell, output activation function, and peephole connection. Memory cell, forgetting gate, and input gate are the key components of LSTM network. The content of memory cell is adjusted by forgetting gate and input gate. When both gates are closed, the content of memory cell will remain unchanged from one time step to the next. The gate structure allows information to be retained across multiple time steps and also allows gradients to be transmitted across multiple time steps, so that the LSTM network overcomes the gradient vanishing problem of RNN [28]. Refer to [28]; Figure 2 shows the structure of LSTM block.

The output and the input of the LSTM block are cyclically connected with the input gate, output gate, and forgetting gate. In Figure 2, the orange rectangle represents the input activation function, and the blue rectangle represents the output activation function; they are usually tan$h$. ● (small black circle) represents the branch point, ⊗ represents dot product, ⊕ is on behalf of all the input, the thin line represents the connection with no weight, the thick line represents the connection with weight, and the dotted line represents the connection with time delay. The three green rectangles represent the input gate, the output gate, and the forgetting gate. They usually use sigmoid activation function to restrict [0,1], where the activation output of 0 means

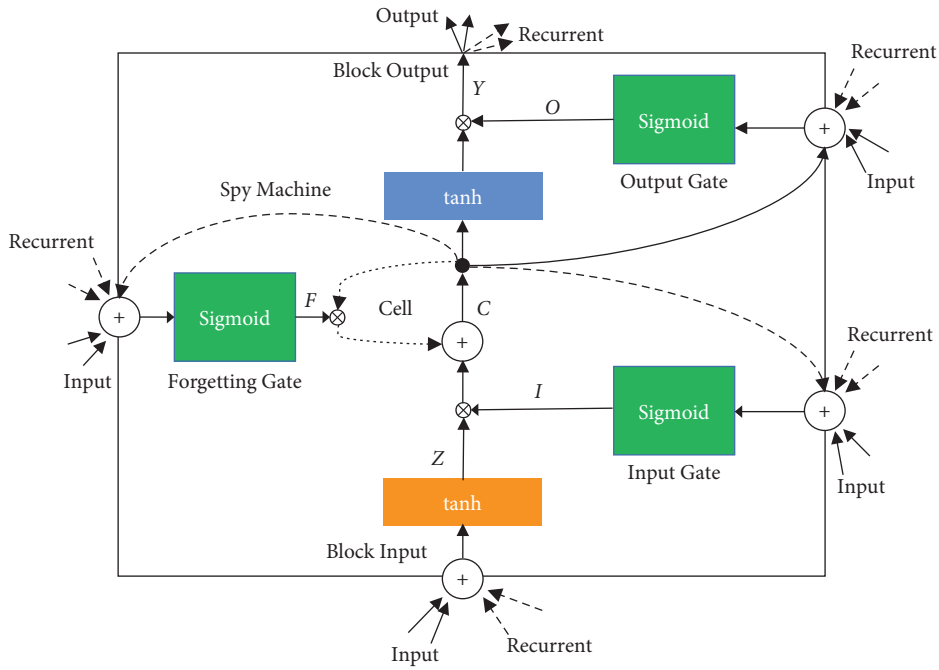


Figure 2: Structure of LSTM block.

"forget everything" and the activation output of 1 means "remember everything."

References [6, 26, 28] give the vector equations of forward transfer of LSTM layer as follows:

$$\begin{aligned} I_t &= \sigma\left(W_I X_t + R_I Y_{t-1} + P_t \otimes C_{t-1} + b_I\right), \\ F_t &= \sigma\left(W_F X_t + R_F Y_{t-1} + P_F \otimes C_{t-1} + b_F\right), \\ C_t &= I_t \otimes Z_t + F_t \otimes C_{t-1}, \\ O_t &= \sigma\left(W_O X_t + R_O Y_{t-1} + P_O \otimes C_t + b_O\right), \\ Y_t &= O_t \otimes \tan h\left(C_t\right), \end{aligned} \tag{2}$$

where P represents the weight vector of the peephole, tan$h$ represents activation function, and $\tan h(x) = ((e^x - e^{-x})/(e^x + e^{-x}))$. Equation (2) represents the input gate, which remembers some current information and determines which value will be updated, protecting the cell from unrelated input events. Equation (2) represents the forgetting gate, which controls how much data is discarded from the current memory state and helps the cell forget the previous memory content. Cell state C_t determines how much information to add or remove from the memory state of the previous time step through sigmoid activation function $\sigma(x) = (1/(1 + e^{-x}))$ and dot product definition layer. Equation (2) represents the output gate, which controls whether the contents of the memory cell are exposed in the LSTM cell. $O_t \otimes \tan h(C_t)$ controls how much memory data will be used in the next phase of the update.

4.2. Prediction Model Construction. We build a deep learning prediction model with two hidden layers of LSTM. The activation function of all layers is tan$h$, which has a more stable gradient and is often used for regression problem [6]. Sigmoid is selected as the gate activation function of the two LSTM hidden layers. The numbers of outputs of the LSTM layers are both 150. The stochastic gradient descent algorithm is used for optimization, and the ADAM algorithm is used for parameter iteration. To solve the overfitting problem of the model, we use the Dropout algorithm developed by Hinton et al. [29] and set the dropout of all LSTM hidden layers as 0.5. Dropout algorithm is a powerful tool to solve the overfitting problem of deep learning models.

Figure 3 shows the basic structure of the prediction model of hotel's actual monthly passenger flow constructed in this paper.

The input layer is specified by the "input_shape" of the first hidden layer of the LSTM model, and its data is a three-dimensional data array. These three dimensions are as follows: sample (a time series), time step (an observation point in the sample represents a time step), and feature (an observation within a time step). In the prediction model in this paper, we set the time steps of the input layer and the feature to 1.

We use Output Layer as the output layer of the prediction model, which has a built-in fully connected Dense Layer. The number of outputs of Output Layer is set to 1, and the function Loss $L1$ is selected as loss function. Loss $l1$ is also called Minimum Absolute Value Deviation and Minimum Absolute Value Error; its purpose is to minimize the sum S of absolute differences between actual value x_i and estimated value $\hat{x}_i$.

$$S = \sum_{i=1}^{n} \left|x_i - \hat{x}_i\right|. \tag{3}$$

5. Experiments

5.1. Data Process. Our original dataset is booking demand data from a resort hotel in Portugal from July 1, 2015, to

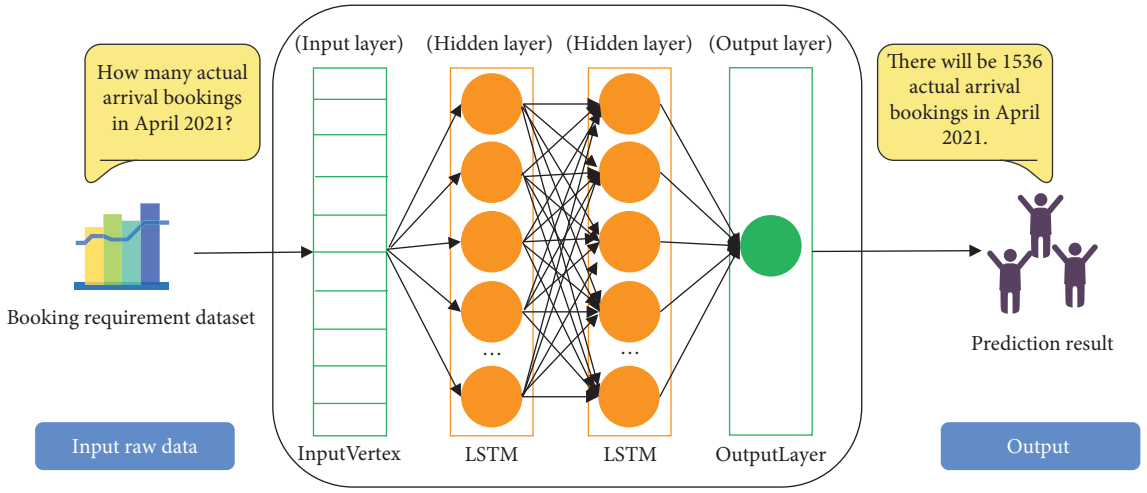


Figure 3: Prediction model structure.

August 31, 2017 [27]. To predict the number of actual arrival bookings in a certain month, firstly, we slice the raw data by time. The size of the time slice is set as one month, starting from July 1, 2015, to July 31, 2015, and ending from August 1, 2017, to August 31, 2017. We calculate the number of all bookings, actual arrival bookings, canceled bookings, and nonarrival bookings in each time slice. Then, we use April 1, 2017, as the time point to split the training set and the testing set; that is, 80% of the dataset are the training set and 20% are the testing set.

5.2. Benchmark Models and Experimental Setup. In order to evaluate the prediction effect of LSTM network under the dataset for hotel booking demand, ANN and SVR models are constructed as the benchmark models. The ANN model adopts a relatively simple single hidden layer network structure. SVR model is one of the most representative nonlinear prediction models. The introduction of SVR model proves that deep learning prediction model in this paper has more advantages than traditional machine learning prediction models in dealing with complex nonlinear data.

We build LSTM, ANN, and SVR models on WEKA 3.8.5 platform of Windows 10 system. In the experiment, we adjust the parameters of all models to a better state to ensure that they can achieve better prediction accuracy. The initial learning rate for all models was set to 0.001. Because the processed dataset is small, we set the batch size of the models to 4. The number of hidden layers in ANN model is 1. The kernel function of SVR model is set as RBF kernel, which contains two important parameters: c and gamma. c is the penalty coefficient, which represents the tolerance to error. When c is larger, overfitting is more likely to occur, and when c is smaller, underfitting is more likely to occur. gamma implicitly determines the distribution of data mapped to a new feature space. When it is larger, the number of support vectors is smaller, which affects the speed of training and testing. Here, we set c to 1.0 and gamma to 0.05.

5.3. Evaluation Metrics. In this paper, we use the three following common model evaluation metrics to evaluate the prediction performance of LSTM, ANN, and SVR. They are mean absolute deviation (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). MAE can well reflect the actual situation of the predicted value error. The larger the MAE is, the worse the effect of the model is; RMSE measures the deviation between the predicted value and the actual value, which is more affected by outliers than the other two evaluation metrics; the value range of MAPE is $[0, +\infty)$. When MAPE is 0%, the model is perfect; when MAPE is more than 100%, the model is poor.

Their calculation equations are as follows:

$$\begin{aligned} \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \hat{x}_i\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2}, \\ \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \times 100\%, \end{aligned} \tag{4}$$

where x_i is the actual value, $\hat{x}_i$ is the predicted value, and n is the number of test samples.

5.4. Result Analysis. We train three models on the training dataset. Table 3 shows the 10-month prediction results of LSTM, ANN, and SVR in the training set, and the best prediction values of each month are shown in red. As can be seen from Table 3, in the 10-month prediction results, LSTM model performs best, SVR performs second best, and ANN performs worst. The optimal prediction results of LSTM, ANN, and SVR are 5 months, 1 month, and 4 months, respectively.

In order to more intuitively compare the prediction effects of the three models on the training set, we draw the prediction curve of each model in Figure 4. The fitting effect

Table 3: Training predictive results of each model.

Time	The number of actual arrival bookings	LSTM	ANN	SVR
July 2016	985	980.4311	990.0974	986.3615
August 2016	1107	1106.7570	1115.8225	1096.7277
September 2016	1068	1069.1288	1106.3889	1069.8033
October 2106	1417	1270.1794	1361.7197	1253.9491
November 2016	1056	1057.2378	1069.8413	1056.4465
December 2016	1056	1056.2752	1076.8899	1055.1749
January 2017	1103	1103.1509	1051.7895	1102.1410
February 2017	1195	1192.0812	1163.5447	1193.9570
March 2017	1164	1161.6975	1229.6771	1164.1725
April 2017	1205	1203.0363	1187.0714	1219.8078

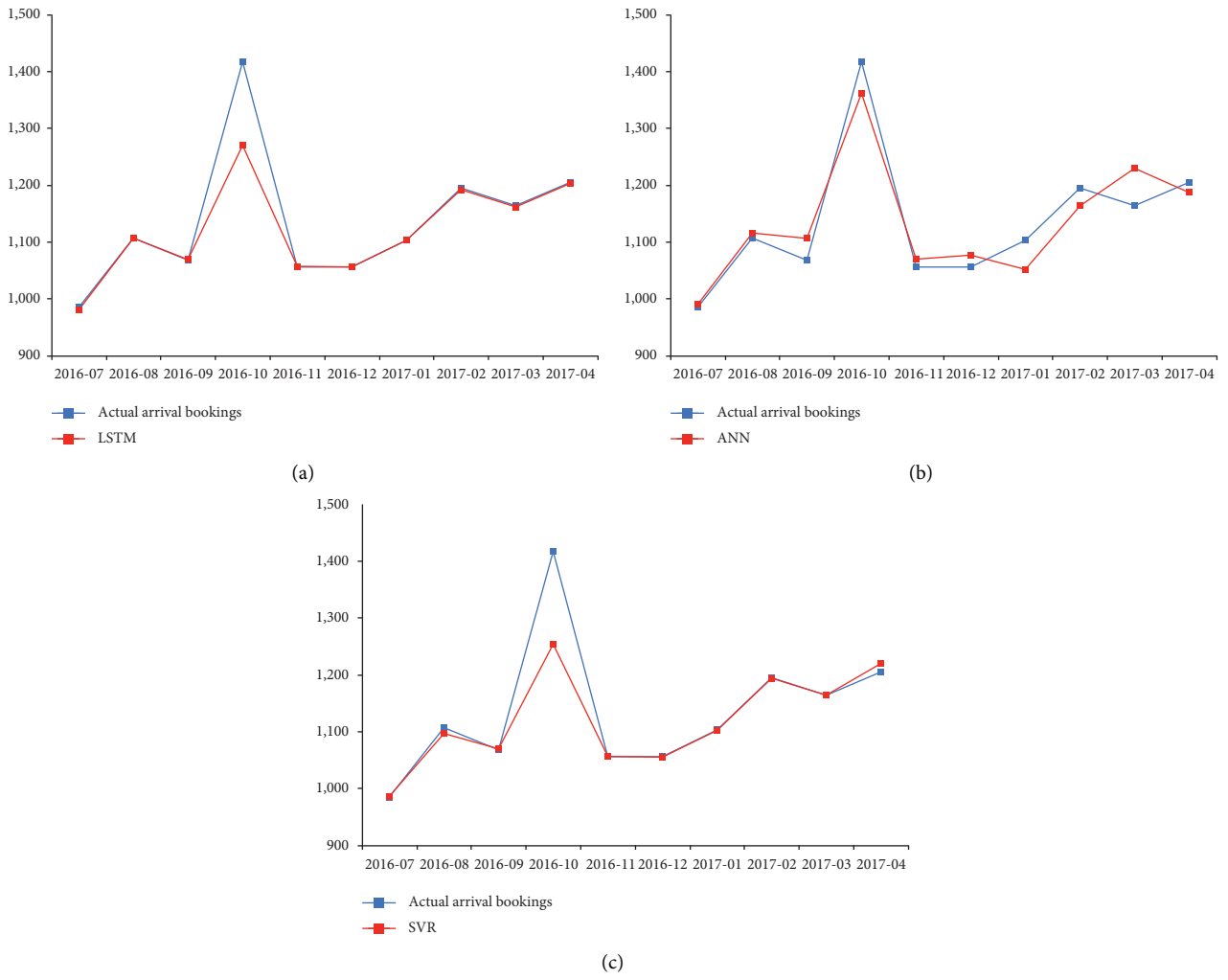


Figure 4: Comparison of the fitting curves of each model in training stage.

of LSTM and SVR model in October 2016 is worse than that of ANN model, but the fitting effect is good in the other nine months. The performance of SVR in August 2016 and April 2017 is slightly worse than that of LSTM model. The fitting effect of ANN model in July, August, and November 2016 is good, but, in January and March 2017, it shows a completely opposite trend to the actual passenger flow.

In general, LSTM model is more consistent with the dynamic characteristics of the actual arrival bookings per month, followed by SVR model, and ANN model is the worst.

In order to evaluate the generalization ability of the model, the optimal structure of the model obtained after training was used as the prediction model for the prediction test. Table 4 shows the 5-month prediction results of LSTM, ANN, and SVR models on the testing set, and the best prediction value of each month is shown in red. As can be seen from Table 4, in the prediction results of 5 months, the

Table 4: Testing predictive results of each model.

Time	Actual arrival bookings	LSTM	ANN	SVR
April 2017	1205	1203.0363	1187.0714	1219.8078
May 2017	1212	1260.8806	1161.1423	1201.0106
June 2017	1045	1088.5516	1074.7438	1143.3589
July 2017	1094	1091.2047	1076.9058	1107.4611
August 2017	1107	1122.6816	1118.6439	1153.5216

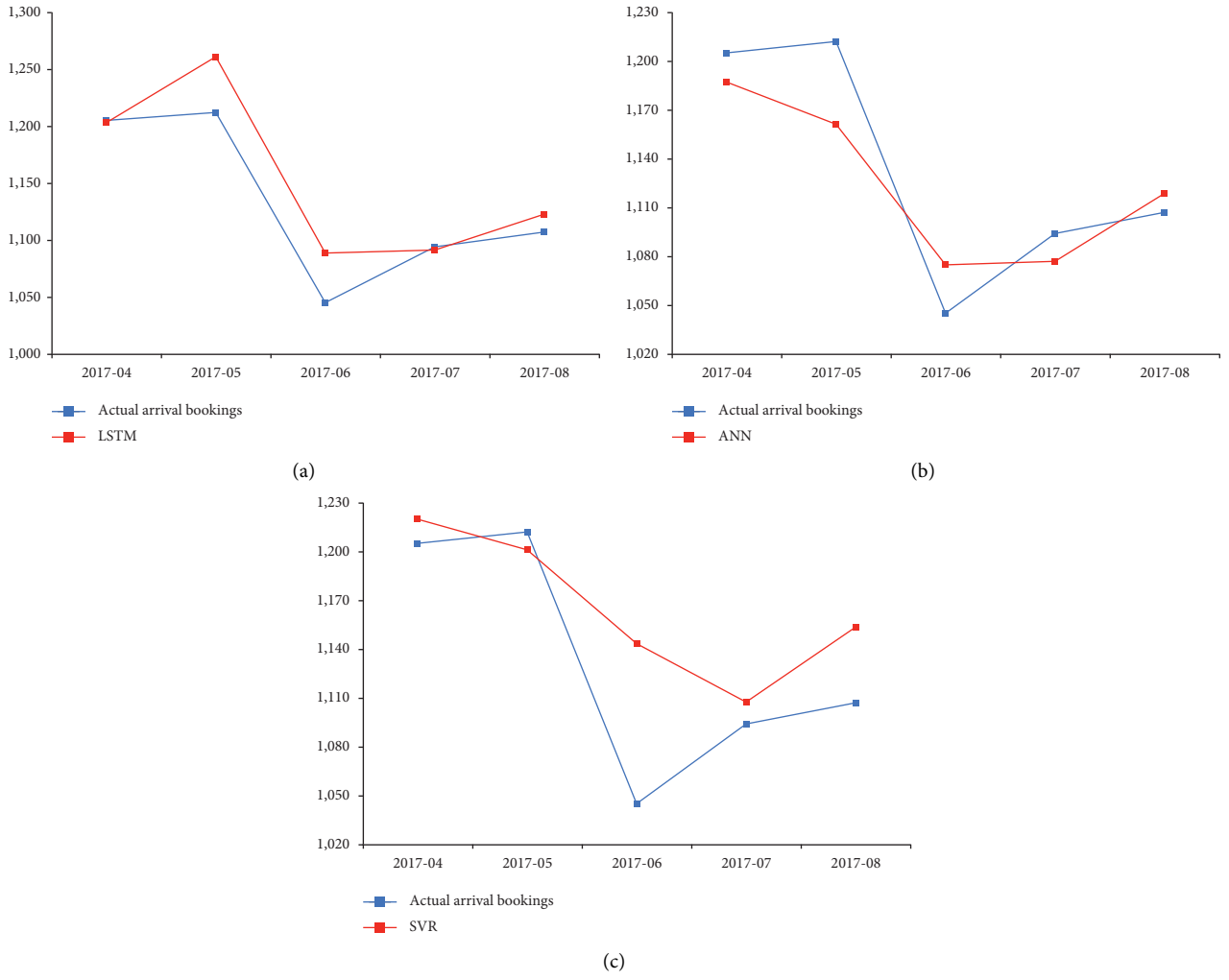


Figure 5: Comparison of the fitting curves of each model in testing stage.

LSTM and ANN models perform the best, while the SVR model performs the worst. The optimal prediction results of LSTM, ANN, and SVR are 2 months, 2 months, and 1 month, respectively.

In order to compare the prediction effects of LSTM, ANN, and SVR on the testing set more intuitively, we draw the prediction curve of each model in Figure 5. In general, the prediction effect of the three models on the testing set is worse than that on the training set. LSTM model has the best fitting effect in April 2017 and July 2017, ANN model has the best fitting effect in August 2017, and SVR model has the best fitting effect in May 2017 and the worst fitting effect in June 2017. As can be seen from Figure 5, from April 2017 to May 2017, the number of actual arrival bookings showed a slight upward trend, while the predicted results of ANN model and SVR model showed a downward trend, which is completely contrary to the reality. From June 2017 to July 2017, the actual arrival bookings showed an obvious upward trend, but the prediction results of the three models in this period showed poor fitting effect, and the prediction results of SVR model are completely opposite to the actual ones.

In general, LSTM and ANN models show better prediction effect than SVR model in testing stage.

In order to further explore the predictive ability of the LSTM model, MAE, RMSE, and MAPE are used to compare and evaluate the three models in the training stage and

Table 5: Evaluation metric scores for each model on the experimental dataset.

Model	MAE		MAPE (%)		RMSE	
	Training	Testing	Training	Testing	Training	Testing
LSTM	17.7385	22.5746	1.2835	2.0071	48.9829	30.1451
ANN	32.2959	25.4518	2.7796	2.2288	38.2780	29.0524
SVR	19.9816	36.8277	1.4491	3.3962	54.4661	49.7191

testing stage. Table 5 records the evaluation metric score of the three models on the experimental dataset, and the best score of each evaluation metric is shown in red.

It can be seen from Table 5 that LSTM model has the best scores on MAE and MAPE but not on RMSE. The overall fitting effect of ANN model on experimental datasets is poor. We can notice that the prediction accuracy of LSTM model in the testing set is close to that in the training set, which indicates that it has the best generalization ability. On the contrary, the performance of SVR model in the testing set is far from that in the training stage.

In the training stage, the score of MAE and MAPE of LSTM model is the lowest, followed by SVR model, and ANN model is the highest. In the testing stage, the scores of MAE and MAPE of LSTM model are still the lowest, followed by ANN model, and SVR model is the highest. As for RMSE, we know that RMSE is very sensitive to outliers, and if there is a predicted value that is very different from the actual value, the score will be very high. In the training stage, ANN model has the lowest score in this metric, the score of LSTM is slightly higher than that of ANN, and SVR is the highest. In the testing stage, the SVR has the highest score on this metric. As can be seen from Figures 4 and 5, in the training stage, the predicted values of LSTM and SVR in October 2016 deviate greatly from the actual values; in the testing stage, the predicted value of SVR in June 2017 deviates greatly from the actual value. Therefore, the RMSE scores of LSTM and SVR are higher than that of ANN model.

Based on the above analysis, LSTM model has a better performance than ANN and SVR models in predicting the number of actual monthly arrival bookings, which can better capture the complex nonlinear characteristics of hotel passenger flow and achieve a better fitting effect.

6. Conclusion

Hotel passenger flow is affected by weather, season, holidays, environment, and other factors, showing characteristics of a complex nonlinear fluctuation. Considering that the traditional demand prediction methods cannot automatically extract the characteristic information from the passenger flow data and cannot deal with the increasing sample data in the practical application, this paper builds a hotel passenger flow prediction model based on deep learning method. Taking a resort hotel in Portugal as an example, we construct an LSTM model with good predictive ability for complex time series to predict the number of actual monthly arrival bookings for the hotel. In order to explore the prediction ability of this model, we construct ANN and SVR as the benchmark models in the experimental stage and compare the prediction effects of the three models on the datasets, with MAE, RMSE, and MAPE as the evaluation metrics. The experimental results show that, compared with the benchmark models, the LSTM model can better simulate the dynamic characteristics of hotel passenger flow and effectively improve the prediction performance and can help hotel managers make more accurate and reasonable pricing decisions and adjust operation mode.

Data Availability

All the data used in this study can be available upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] M. Zhang, *Hotel Occupancy Prediction Based on EEMD-ARIMA*, Shaanxi Normal University, Xi'an, China, 2016.

[2] C. Xu, "Hotel revenue management: research overview and prospects," *Journal of Management Sciences in China*, vol. 29, no. 6, pp. 72–78, 2003.

[3] W. Wu, "Prediction of hotel occupancy based on support vector regression," *New Technology & New Products of China*, vol. 3, no. 361, pp. 125-126, 2018.

[4] L. R. Weatherford and S. E. Kimes, "A comparison of forecasting methods for hotel revenue management," *International Journal of Predicting*, vol. 19, no. 3, pp. 401–415, 2003.

[5] J. Yu, *Model Established and Analysis of Passenger Flow Forecast in Hotel Management Decision Support System*, Dalian University of Technology, Dalian, China, 2008.

[6] B. Zhang, Y. Pu, Y. Wang, and J. Li, "Forecasting hotel accommodation demand based on LSTM model incorporating internet search index," *Sustainability*, vol. 11, no. 17, p. 4708, 2019.

[7] S. Y. Ji, "Hotel occupancy prediction based on singular spectrum analysis," *Information Recording Materials*, vol. 19, no. 3, pp. 49–51, 2018.

[8] Y.-M. Zhang, "Research summary of demand forecasting based on hotel revenue management," *Sci-Tech Innovation & Productivity*, vol. 26, no. 7, pp. 7–12, 2019.

[9] B.-R. Zhang, "Research on tourism demand forecast based on consumer search within internet environment," University of International Business and Economics, Beijing, China, 2017.

[10] W. P. Andrew, D. A. Cranage, and C. K. Lee, "Forecasting hotel occupancy rates with time series models: an empirical analysis," *Hospitality Research Journal*, vol. 14, no. 2, 1990.

[11] P. E. Pfeifer and S. E. Bodily, "A test of space-time ARMA modelling and forecasting of hotel data," *Journal of Forecasting*, vol. 9, no. 3, pp. 255–272, 1990.

[12] B. Pan, D. C. Wu, and H. Song, "Forecasting hotel room demand using search engine data," *Journal of Hospitality and Tourism Technology*, vol. 3, no. 3, pp. 196–210, 2012.

[13] J.-J. Yang, "Design of resort hotel passenger flow prediction system based on logistic model," *Journal of Changchun Normal University*, vol. 39, no. 8, pp. 54–59, 2020.

[14] X. Sun, *Cruise Line Revenue Management: Demand Forecasting and Revenue Optimization*, Shanghai Jiao Tong University, Shanghai, China, 2011.

[15] C. Goh and R. Law, "Modeling and forecasting tourism demand for arrivals with stochastic non-stationary seasonality and intervention," *Tourism Management*, vol. 23, no. 5, pp. 499–510, 2002.
[16] C. Goh, "Exploring impact of climate on tourism demand," *Annals of Tourism Research*, vol. 39, no. 4, pp. 1859–1883, 2012.
[17] J.-G. Choi, "Developing an economic indicator system (a forecasting technique) for the hotel industry," *International Journal of Hospitality Management*, vol. 22, no. 2, pp. 147–159, 2003.
[18] W.-C. Hong, Y. Dong, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic genetic algorithms for tourism demand forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 1881–1890, 2011.
[19] A. Rashad, S. Sara, and A. Rafig, "Development of fuzzy time series model for hotel occupancy forecasting," *Sustainability*, vol. 11, no. 3, p. 793, 2019.
[20] S. Sun, Y. Wei, K.-L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," *Tourism Management*, vol. 70, pp. 1–10, 2019.
[21] L. E. I Ke-wei and Y. Sheng, "Forecast of inbound tourists to China based on BP neural network and ARIMA combined model," *Tourism Tribune*, vol. 22, no. 4, pp. 20–25, 2007.
[22] W. Gong and J. Huang, "A demand forecast model based on the gray theory and exponential smoothing method," *Statistics & Decision*, vol. 33, no. 1, pp. 72–76, 2017.
[23] B.-R. Zhang, S.-L. Liu, C.-F. Zhang, and Y.-L. Pu, *Forecasting hotel occupancy rate based on consumer search within network environment*, Statistics & Information Forum, vol. 33, no. 3, 2018.
[24] Y. Chang and C. Tsai, "Apply deep learning neural network to predict number of tourists," in *Proceedings of the 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 259–264, Taipei, Taiwan, 2017.
[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] R. Law, G. Li, D. K. C. Fong, and X. Han, "Tourism demand forecasting: a deep learning approach," *Annals of Tourism Research*, vol. 75, pp. 410–423, 2019.
[27] N. Antonio, A. De Almeida, and L. Nunes, "Hotel booking demand datasets," *Data in brief*, vol. 22, 2019.
[28] J. Patterson and A. Gibson, *Deeplearning: A Practitioner's Approach by Josh Patterson and Adam Permissions*, O'Reilly Media, Inc., Sebastopol, CA, USA, 2017.
[29] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

Hindawi
Complexity
Volume 2021, Article ID 5556899, 10 pages
https://doi.org/10.1155/2021/5556899

Research Article

Commodity Image Classification Based on Improved Bag-of-Visual-Words Model

Huadong Sun,[1,2,3] **Xu Zhang**,[1,2] **Xiaowei Han**,[1,2] **Xuesong Jin**,[1,2] **and Zhijie Zhao**[1,2,3]

[1]*School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China*
[2]*Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin 150028, China*
[3]*North-East Asia Service Outsourcing Research Centre, Harbin University of Commerce, Harbin 150028, China*

Correspondence should be addressed to Huadong Sun; kof97_sun@163.com

Received 4 February 2021; Revised 1 March 2021; Accepted 4 March 2021; Published 17 March 2021

Academic Editor: Abd E.I.-Baset Hassanien

With the increasing scale of e-commerce, the complexity of image content makes commodity image classification face great challenges. Image feature extraction often determines the quality of the final classification results. At present, the image feature extraction part mainly includes the underlying visual feature and the intermediate semantic feature. The intermediate semantics of the image acts as a bridge between the underlying features and the advanced semantics of the image, which can make up for the semantic gap to a certain extent and has strong robustness. As a typical intermediate semantic representation method, the bag-of-visual-words (BoVW) model has received extensive attention in image classification. However, the traditional BoVW model loses the location information of local features, and its local feature descriptors mainly focus on the texture shape information of local regions but lack the expression of color information. Therefore, in this paper, the improved bag-of-visual-words model is presented, which contains three aspects of improvement: (1) multiscale local region extraction; (2) local feature description by speeded up robust features (SURF) and color vector angle histogram (CVAH); and (3) diagonal concentric rectangular pattern. Experimental results show that the three aspects of improvement to the BoVW model are complementary, while compared with the traditional BoVW and the BoVW adopting SURF + SPM, the classification accuracy of the improved BoVW is increased by 3.60% and 2.33%, respectively.

1. Introduction

With the increasing of e-commerce, online shopping has become the main way for the public to buy goods. In order to provide a better shopping experience for users to quickly browse and search for goods, a good commodity image classification system also highlights its importance. In the past, the image classification method based on manual tagging has not met the actual needs [1]. How to use image processing, computer vision, pattern recognition, and machine learning to realize the classification of commodity images has great research and commercial value.

The process of image classification includes feature extraction, classifier training, and classification. In the mature development of classifier, image feature extraction often determines the final effect. Whether in the field of image recognition, image retrieval, image classification, or image ranking, feature extraction is very important [2, 3]. At present, the image feature extraction mainly includes the underlying visual feature and the intermediate semantic feature [4, 5]. The underlying visual features mainly refer to the color, texture, and shape of the image. By extracting the underlying features, the high semantics of the image can be inferred and the image classification can be realized. The commonly used underlying visual features are as follows.

Color feature is one of the most commonly used underlying features, which has the advantages of simple extraction, rotation, translation and scale invariance, and clear physical meaning. Color is also an important basis for commodity image classification and retrieval. Different types

of commodity images show great differences in color proportional distribution and color spatial distribution. The color histogram algorithm is proposed for the first time by Swain and Ballard [6], which is an intuitive expression of image content. By counting the frequency of different color pixels in the image, the composition of color is reflected. Strieker and Orengo et al. [7] proposed the color moment, which does not need vector quantization and directly accumulates statistics on each channel, and then only nine values are needed to describe the feature information. However, the recognition accuracy of this method is low, so it is necessary to combine other methods to characterize the image. Naushad et al. [8] constructed three probability histograms for each color component, which were then divided into numbers of several valid intervals and calculated statistics such as standard deviation, skewness, and kurtosis from each interval which were used as image color features.

Texture feature is based on the gray level statistics, describing the smoothness, roughness, and appearance law, reflecting the structure information of the image and the spatial distribution of the gray level. Texture features, as inherent properties of object surfaces, are of great significance in the classification of commodity images. Ojala et al. proposed LBP operator [9, 10] for texture classification. This method has small computational complexity and multiscale and rotation invariant properties and is widely used in texture retrieval. With the depth study of LBP rotation invariance, other forms, such as LBP variance and global matching [11], complete model [12] of LBP, joint distribution [13] of simulated Gaussian mixed local patterns, and modified LBP [14], have been effectively applied in texture classification.

Shape can distinguish the different objects in the image more intuitively; shape features are usually related to the target which the user is interested. Shanmugavadivu et al. [15] used fuzzy-object-shape to capture the shape of the object, improving the accuracy of boundary information, and provided an approximation measure of the object by a conventional shape. Wu et al. [16] proposed a new algorithm to calculate the rotation invariant of Tchebichef moment, and translation and scale invariance of Tchebichef moments are achieved by prealigning the image into a standard image. The proposed descriptor is compared with radial Tchebichef moment, and two kinds of Zernike moment experimental results show that the proposed shape features are robust to deformations generated by image shape rotation and scaling. Sokic et al. [17] proposed an improved Fourier descriptor method, which is capable of extracting Fourier descriptors in condition of translation, scaling, rotation, and starting point changes.

Commodity images are rich in color, shape, and texture information. Such underlying features provide simple representation of images based on physical level. Once there is a large change in the class or there is a significant background interference, the classification accuracy will be reduced. The intermediate semantics of the image acts as a bridge between the underlying features and the high semantics of the image, which can make up for the semantic gap to a certain extent robustness. As a typical intermediate semantic representation method, the bag-of-visual-words (BoVW) model [18] has received extensive attention in image classification. However, the traditional BoVW loses the location information of local features, and its local feature descriptors always lack the expression of color information.

In recent years, deep learning has become a research hotspot in the field of machine learning and artificial intelligence. As a feature learning method, deep learning has achieved good results in image classification. The AlexNet [19] model is the beginning of CNN widespread concern, followed by many innovative classification structure models, such as GoogLeNet [20], Inception v3 [21], ResNet [22], and so on. However, the structure of the CNN is complex, and it takes a lot of time in the calculation process, and the network lacks the necessary interpretability.

In this paper, the improved BoVW for the commodity image classification according to the shortcomings of traditional BoVW and the characteristics of commodity images is studied, which explore a more reasonable local feature description and add a description of location information to the model.

2. Principle of Bag-of-Visual-Words Model

The bag-of-visual-words (BoVW) model is a natural extension of the bag-of-words (BoW) model from natural language processing field to image processing field [18]. The principle of BoVW is described as follows: dividing the image into small pieces and then clustering similar chunks into visual words, BoVW counts the frequency of these visual words in the image, represented in the form of a histogram. Generally, image local features are used to compare words in the BoW model, such as SIFT [23] and SURF [24]. Since the importance of each visual word to different categories of images is different; hence, image classification can be performed by BoVW combined classifier (such as SVM). The schematic diagram of BoVW is illustrated in Figure 1.

By using the BoVW model to represent the image and obtain the global histogram representation of the image, there are five steps:

Step 1. Automatically detect the key points of the image and search for local regions.

Step 2. Feature extraction of local regions: According to the specific application considerations, the uniqueness of the features, the complexity of the extraction algorithm, and the effect of the selection features, the local feature extraction algorithm is used to extract local features from images.

Step 3. Visual dictionary construction: Generally speaking, a part of the image from different categories is selected from the image library to form the training image set, and its local features are extracted. Then, all the local feature vectors of the training image are defined as visual words by proper redundancy processing. The usual processing method is to cluster all the local feature vectors of the training image and define the cluster

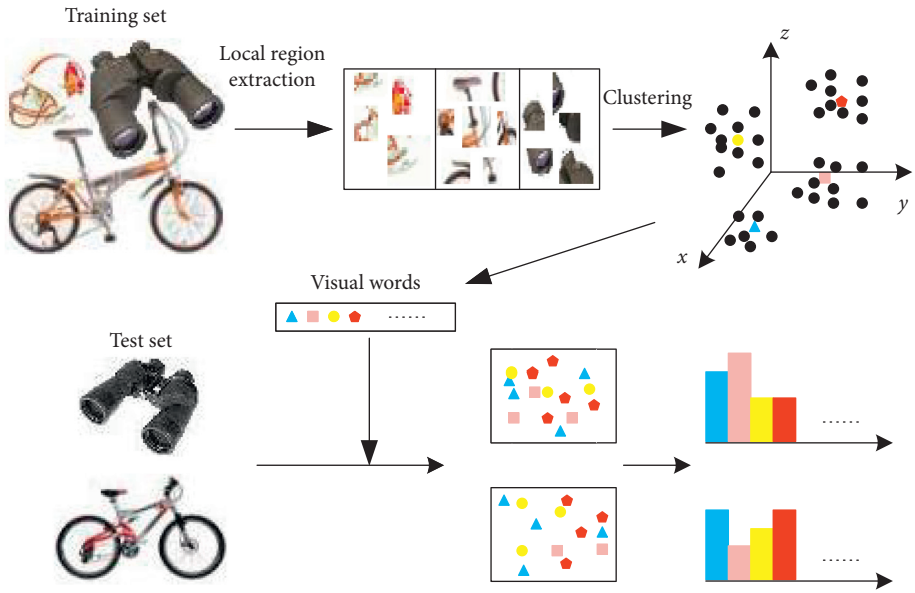


Figure 1: Schematic diagram of BoVW.

centre as a visual word. All visual words form a visual dictionary for histogram representation of images.

Step 4. Feature vector encoding: The BoVW model uses vector quantization technology to encode feature extraction of local regions. The result of vector quantization is to quantify the local feature vector of the image into the visual word which is the most similar. The vector quantization process is actually a search process. Usually, the nearest neighbor search algorithm is adopted to search for the most matching visual words with the local feature vectors of local regions.

Step 5. Use a visual word histogram to represent the image: After all the local feature vectors of an image are quantized, the frequency of each visual word in the visual dictionary can be counted, and a histogram about the visual word can be obtained. Its essence is the global statistical result of quantization coding obtained in the previous step. It is a numerical vector composed of visual word index order. This vector is the final representation of the image.

3. Improved Bag-of-Visual-Words Model

The traditional BoVW loses the location information of local features, and its local feature descriptors mainly focus on the texture or shape information of local regions but lack the expression of color information. In this paper, the BoVW model will be improved. Technical details include (1) multiscale local region extraction; (2) local feature description by SURF + CVAH; (3) dictionary generation and feature coding; and (4) diagonal concentric rectangular pattern.

3.1. Multiscale Local Region Extraction (MLRE). Multiscale local region extraction is the first step in BoVW, including multiscale key point extraction, location mapping, and region division. The methods to extract key points are Harris operator, Fast operator, and SUSAN operator. These algorithms are very strict in the selection of key points, so they are very suitable for image matching. However, when applied to BoVW, too strict position selection will lead that some effective regions cannot be extracted sufficiently, resulting in insufficient information extraction of local regions. Hence, the multiscale feature of wavelet transform is used to solve this problem, and the local region extraction algorithm suitable for the BoVW model is explored.

3.1.1. Selection of Multiscale Key Points. As we know, four subimages at the current layer can be obtained after wavelet decomposition of two-dimensional images including a rough sub-band (low frequency component); and there are detailed sub-bands (high frequency component) in three directions: horizontal, vertical, and diagonal. Each rough sub-band can continue to do the next level of decomposition.

Here, wavelet decomposition is used to carry on the multiscale analysis to the image, which realizes the multiscale key point extraction, as is shown in Figure 2. The process is as follows:

(1) In order to obtain more key points, the commercial image grayscale value and double up-sampling processing are adopted.

(2) Multilayer wavelet decomposition to up-sampled image is carried on, and the number of decomposition layers is 3.

(3) High frequency sub-bands' coefficients are normalized, and candidates are selected according to coefficients in condition that the coefficients of three high frequency channels are greater than 0.1 in the same coordinates.

(4) Nonmaximum suppression is carried on to all candidates, and then the key points in the

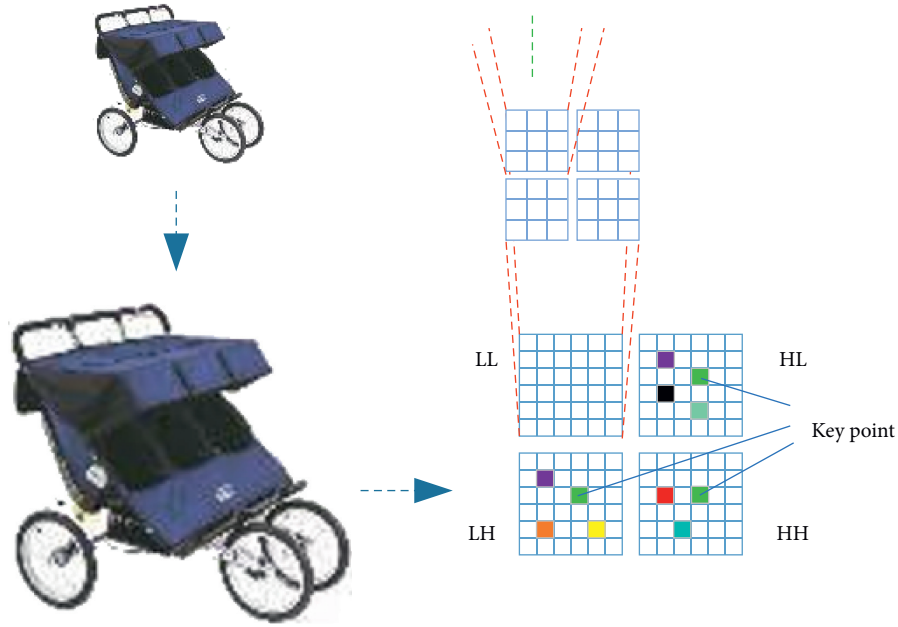


Figure 2: Schematic diagram of multiscale key point extraction.

corresponding scale are obtained. The detailed process is as follows. The matrix block of size 5×5 is delimited with the centre of candidate point. The value of each position of matrix block is the sum of the response values of three wavelet high frequency sub-bands' coefficients in the same coordinates. If the value of the candidate point is the maximum value of the region, it is retained as the key point.

3.1.2. Location Mapping and Local Region Extraction. Because the feature points are detected at different wavelet layers and each layer is carried on, the image is reduced by a quarter. In order to describe the local region features with more abundant information, it is necessary to map the coordinates to the original map. Location mapping relationship is described as

$$\begin{cases} X = x^L \cdot \left(2^{(L-1)} - k \cdot 2^{(L-2)}\right), \\ Y = y^L \cdot \left(2^{(L-1)} - k \cdot 2^{(L-2)}\right), \end{cases} \quad k = \begin{cases} 0, & L = 1, \\ 1, & L > 1, \end{cases} \tag{1}$$

where L is the number of wavelet decomposition layers, x and y are the position coordinates of key point detected on the scale, and X and Y are the coordinate positions corresponding to the original image.

It is worth noting that in image matching, the region size of key points needs to be calculated accurately, but in the BoVW model, the scale parameter is only a measure of the local region size. Therefore, it is not necessary to calculate the accurate scale by interpolation. The scale parameter is determined according to the following formula:

$$\sigma = 1.6 * 2^{L-1}, \tag{2}$$

where L is the number of wavelet decomposition layers, and 1.6 is the initial scale. Then, we can define local region as the square with length 20σ, whose centre is key point. However, another need to be determined is the main direction of local region, which will be introduced in detail in Section 3.2.1.

3.2. Local Feature Description by SURF + CVAH. The local feature description requires the balanced uniqueness of the balance. Uniqueness is so strong to distinguish between the two features which originally belong to the same visual word, while uniqueness is so weak to cause the visually distinct regions to be regarded as the same in the visual dictionary, which is obviously inappropriate. In addition to the uniqueness or differentiability, the local feature description generally requires the following properties: rotation invariance, illumination invariance, and noise resistance.

Color, texture, and shape, as three underlying features, play different roles in describing images. In the commodity image, the color of commodity is changeable, and even the same product has many colors, which leads to the fact that some color descriptors, such as color histogram and color moment, are not suitable. On the other hand, the texture of commodity image is relatively single, and there are many similar texture features between different commodities. Therefore, it is not enough to describe the commodity image by texture to satisfy the balance of descriptor's uniqueness. Therefore, this paper uses the series of texture and color features to solve this problem.

3.2.1. Speeded Up Robust Features (SURF). SURF can achieve a good local region texture description, which can be divided into two processes. (1) Main direction confirmation: in order to achieve the rotation invariance of features, we need to confirm the main direction of feature points. The direction is chosen around the feature point in a small area of scale. (2) Region feature generation: the feature region is

rotated to the main direction, and the texture information of local region is extracted.

In the SURF algorithm, the main direction is determined by the response value of Haar wavelet which is applied to multiple subdomains in the circular scale region. Specifically, the circular scale region with a radius of 6σ is taken as the local region of key point, and the sum of the horizontal and vertical Haar wavelet response values of all pixel points in the 60° sector region is counted. Then, the sector rotates at an interval of 11.46° and counts the sum of wavelet response values in the region after rotation. The wavelet response values of all sampling directions are obtained by pushing it until the sector region is rotated whole circle. The direction of the sector with the largest response value is selected as the main direction of key point. The schematic diagram is shown in Figure 3.

After the main direction selected, the coordinate system of image is rotated in the main direction so that the coordinate axis is consistent with the main direction. Then, $20\sigma \times 20\sigma$ rectangular areas are taken around the key point and this region is divided into 16 rectangular subregion blocks whose size is $5\sigma \times 5\sigma$. Next, calculating the Haar wavelet response value of subregion blocks, we can obtain the sum of response values in horizontal direction $\sum dx$, the sum of response values in vertical direction $\sum dy$, the sum of the absolute values in horizontal direction $\sum |dx|$, and the sum of the absolute values in vertical direction $\sum |dy|$. So, we can get the feature vector $(\sum dx, \sum dy, \sum |dx|, \sum |dx|)^T$ from one subregion block. Finally, after all feature vectors of 16 subregion blocks are concatenated in the same order, the SURF vector whose size is 64×1 can be obtained. The schematic diagram is illustrated in Figure 4.

3.2.2. Color Vector Angle Histogram (CVAH). A variety of methods are used to measure the difference between two kinds of colors in RGB color space, and the most commonly used distance measurement method is Euclidean distance. The Euclidean distance calculation method is simple and easy, whose characteristic is rotation invariance. However, the RGB color model is not uniform space, and its visual differences can be hardly are reflected by Euclidean distance, which exposes the shortcoming of Euclidean distance. Therefore, using angles to measure color differences is a good choice. In the RGB space, CVA represents the angle between the RGB color vector of two adjacent pixels, as shown in Figure 5.

The formula for calculating the color vector angle is as follows:

$$\theta = \arccos\left(\frac{r_1 r_2 + g_1 g_2 + b_1 b_2}{\sqrt{r_1^2 + g_1^2 + b_1^2}\sqrt{r_2^2 + g_2^2 + b_2^2}}\right), \quad (3)$$

where (r_1, g_1, b_1) is the color vector of a pixel in RGB space, (r_2, g_2, b_2) is that of adjacent pixels, and θ is the color vector angle between the two pixels.

The implementation process is as follows. In the neighborhood centered on the key points, according to formula (3), calculate the color vector angle between each pixel point and the key point in the local region. Then, quantify the color vector angle uniformly, count the number of pixels in each interval segment, and obtain the color vector angle histogram (CVAH). The color vector angle reflects the color difference of each pixel point in the local region relative to the central key point. Generally speaking, the color vector angle is between 0° and 90°. Here, the quantization step is chosen as 0.5°, so that the dimension of color vector angle histogram is 180×1.

The final local region feature description vector can be obtained by splicing the 64-dimensional SURF vector with the 180-dimensional CVAH, which is 244-dimensional and can effectively describe the color, shape, and texture information of the local region.

3.3. Dictionary Generation and Feature Coding. After local features extraction, each feature represents a local region, and visual words can be obtained by the clustering algorithm. This is a typical dictionary generation and coding problem. The detail is as follows.

T_i is the local feature vector, whose size is $D \times 1$; that is, $T_i \in R^{D\times 1}$. Obviously, if the feature descriptor introduced in Section 3.2 is used, T_i is made of SURF and CVAH splicing, that is; $D = 244$. Suppose there are N local feature vectors corresponding to all local regions from all training images, the set composed by all local feature vectors can be written as

$$T = [T_1, T_2, \ldots, T_N], \quad T \in R^{D\times N}. \quad (4)$$

Then, K-means quantization can be described by the following formula:

$$\begin{aligned} &\min_{U,V} \quad \sum_{i=1}^{N} \left\| T_i - V \cdot u_i \right\|^2, \\ &\text{s.t.} \quad \text{Card}(u_i) = 1, \quad \left\| u_i \right\|_1 = 1, \; u_i \geq 0. \end{aligned} \quad (5)$$

In formula (5), V is codebook (dictionary), which contains K visual words (cluster centers), that is,

$$V = [V_1, V_2, \ldots, V_K], \quad V \in R^{D\times K}, \quad (6)$$

where column vector $V_i (i = 1, 2, \ldots, K)$ is visual word and $V_i \in R^{D\times 1}$.

In formula (5), u_i is the code which is generated by using codebook V to encode local feature vector T_i and u_i is a column vector of K dimension, $u_i \in R^{K\times 1}$, while $\text{Card}(u_i) = 1$ and $\|u_i\|_1 = 1$ can ensure that one element of u_i can be taken as 1 and the other elements of u_i are all 0. The encoding matrix of all local feature vectors in the training set can be described as

$$U = [u_1, u_2, \ldots, u_N], \quad U \in R^{K\times N}. \quad (7)$$

In the quantization process, the BoVW model assigns a local feature vector to a unique visual word closest to it. The index of the only nonzero element in the code u_i indicates the cluster centre to which the local feature vector T_i belongs. Such an optimization problem can be transformed

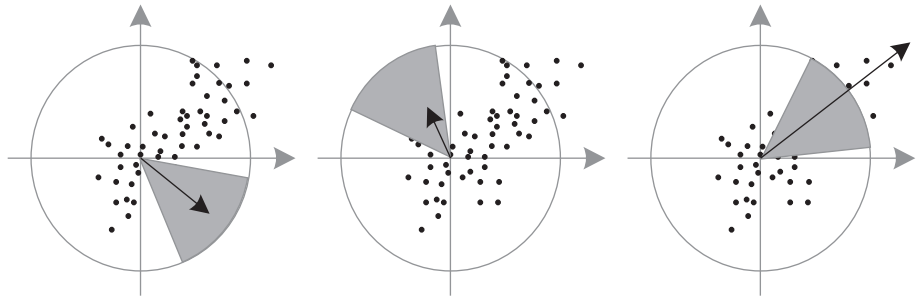


Figure 3: Main direction selection.

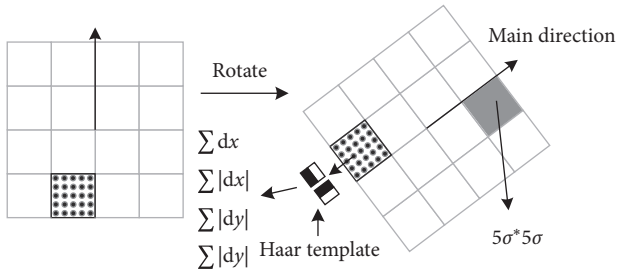


Figure 4: Region feature generation.

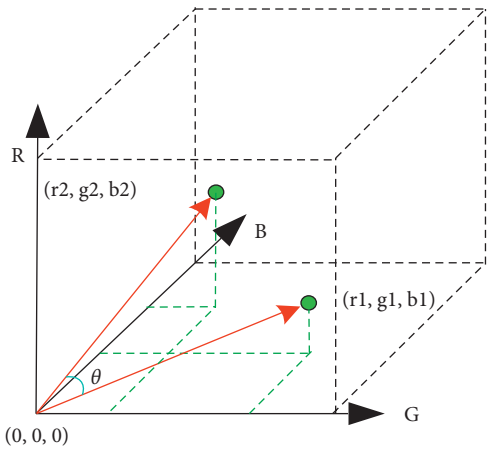


Figure 5: Color vector angle.

into a matrix decomposition problem of the encoding matrix U.

To an image, the final feature is the number of each visual words appearing in the whole dictionary. After normalization, the frequency of each visual word can be obtained. Suppose the local feature vector of an image corresponds to the first M column vectors of the matrix T, then the visual word histogram of this image is counted as

$$H = \frac{1}{M}\sum_{i=1}^{M} u_i. \tag{8}$$

The visual word histogram is a K-dimensional column vector, which can be used as the middle semantic expression of the image. The visual word histograms of all pictures in the training set can be obtained, and the classifier can be trained to guide the classification of the test images.

3.4. Diagonal Concentric Rectangular Pattern (DCRP). The traditional BoVW model discards the sequential relationship of feature descriptors, and the global histogram is completely missing the information of local features, which limits its representation ability. Spatial pyramid matching (SPM) was proposed to make up for its deficiency. In SPM, considering the spatial information, the image is divided into several blocks (sub-blocks), and the features of each sub-block are counted separately. Finally, the features of all blocks are spliced together to form the complete features, which are the distribution of image feature at different resolutions, and can obtain the local information of the image. During the partitioning process, SPM adopts a multiscale method to make its structure present a hierarchical pyramid shape. The SPM model is suitable for most scenes, but it also lacks particularity, which is not the best representation in the case of spatial distribution with characteristic law. Therefore, we propose diagonal concentric rectangular pattern (DCRP) which is suitable for describing the spatial distribution characteristics of commodity images.

It is necessary to analyze the characteristics of commodity images in e-commerce platforms. Firstly, commodity images of the same e-commerce platform always have a uniform resolution; that is, the row and column size is equal or close to equal. Secondly, the background of commodity images is always monotonous, which can be clearly divided into foreground region with a large amount of commodity information and monochromatic background region with noise points. Thirdly, the commodity target position in commodity images is generally in the middle.

Taking into account the spatial distribution characteristics of commodity images and drawing lessons from SPM ideas, we propose diagonal concentric rectangular pattern (DCRP), as shown in Figure 6. The scale definition of DCRP is slightly different from that of SPM. Scale 0 refers to the whole commodity image. Scale 1 refers to the central square region composed of four small squares, and the surrounding region composed of four trapezoids. Scale 2 refers to the 4 small squares and 4 trapezoids. Then, there are total $1+2+8=11$ blocks in all scales of DCRP. The final representation vector of the image is 11K dimension, which is less than 21K dimensions of SPM, which greatly reduces the computation. DCRP effectively introduces the location information of local features in commodity images, which improves the representation ability of BoVW.

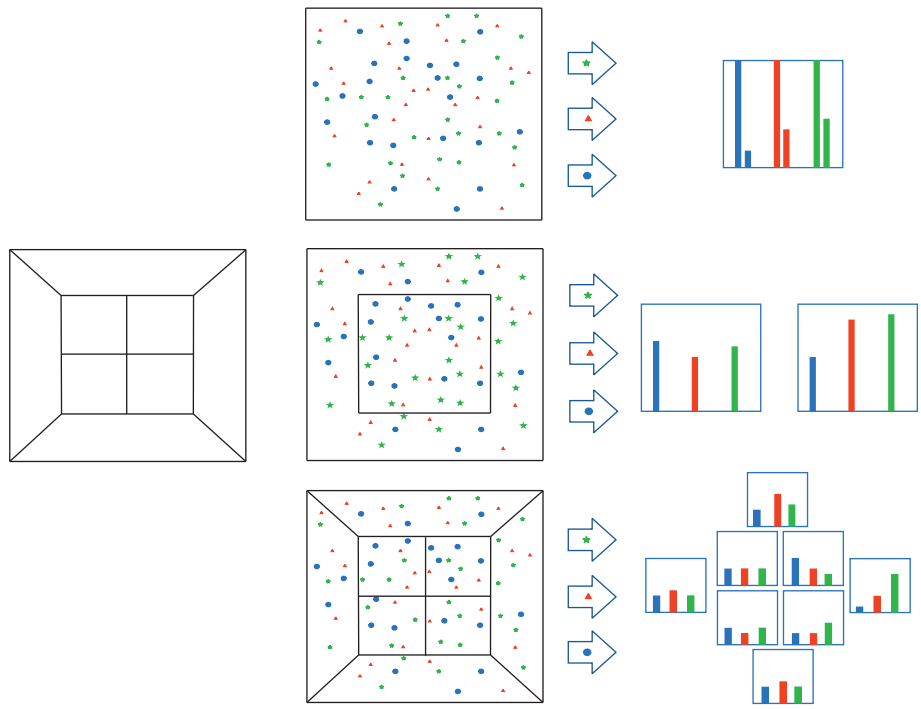


Figure 6: Schematic diagram of diagonal concentric rectangular pattern.

4. Experimental Results and Analysis

Experimental dataset is the Microsoft PI100, which has 100 kinds of commodity images, and each class has 100 images, a total of 10000 images. In the experiment, we select 10 kinds of commodity images. For each kind of images, 30 pictures are selected as the training set and 70 as the test set. SVM (support vector machine) is adopted to classify. Because the clustering centers obtained by the K-means algorithm are not completely consistent for each run, the classification accuracy is calculated by averaging the results of 10 runs. Experimental simulation environment is as follows: CPU: Intel Core i5-5200U; memory: 4G; simulation platform: Matlab R2018.

4.1. Effects of Multiscale Local Region Extraction on Classification. In Figure 7, the red curve is the result of BoVW with SIFT, the green is that of SURF, and the black one is that of SURF with the proposed multiscale local region extraction (for short, MLRE) algorithm. It can be seen clearly, with the increase in the number of visual words in BoVW, the classification effect is incremental. When the number of visual words tends to 1000, classification accuracy tends to be close to the limit. When the number of visual words is 1000, the accuracy of SURF (MLRE) is 87.12%, 1%, and 15.1% higher than that of SURF and SIFT, respectively. It explains that MLRE proposed here is effective to commodity image classification.

4.2. Comparison of Different Descriptors of Local Region. Figure 8 shows the contrast of different local area descriptors, where the classification effect is incremental with

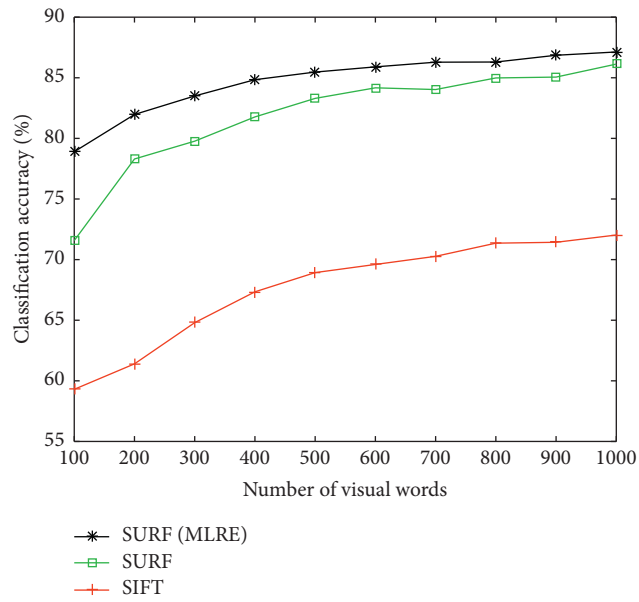


Figure 7: Effects of multiscale local region extraction on classification.

the increase in the number of visual words in the BoVW model. When the number of word packets is 900, without MLRE, the accuracy of BoVW using SURF as local feature description is 85.06%, while that of the BoVW model adopting SURF + CVAH is 86.80%, and there is 1.74% increase. On the other hand, when MLRE is used, the accuracy of BoVW using SURF is 86.86, while that of the BoVW model adopting SURF + CVAH is 88.69%, increased by 1.83%. As can be seen from the above results, adding CVAH

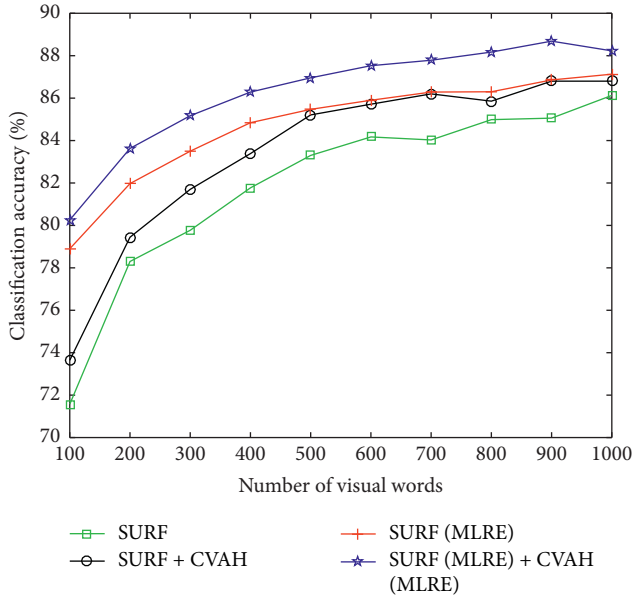


Figure 8: Comparison of different descriptors of local region.

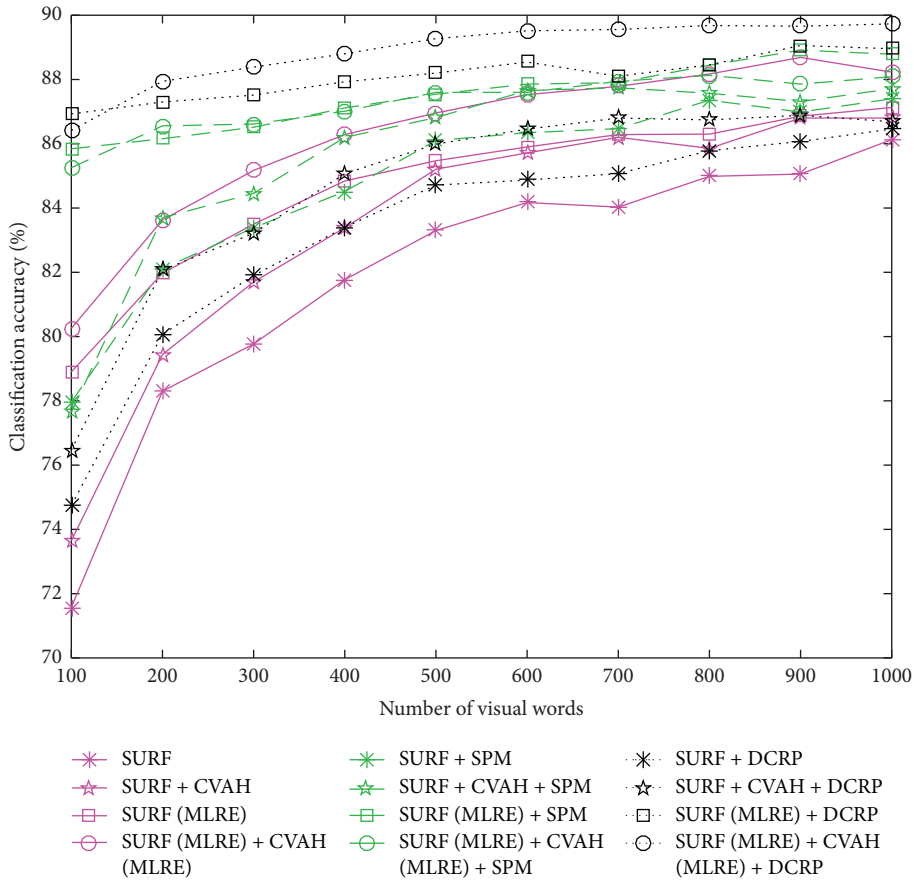


Figure 9: Effects of different spatial patterns on classification.

to local feature description is beneficial to the improvement of classification accuracy. And it can also be seen that the classification accuracy of MLRE and CVAH is 3.63% higher than that of SURF alone.

4.3. Effects of Different Spatial Patterns on Classification. Figure 9 shows the effects of different patterns on classification, where the four pink lines represent the classification accuracy of tradition BoVW without any spatial pattern, the

Table 1: Comparison with other image classification methods on Microsoft PI100 dataset.

Method	Classification accuracy (%)
LBP + SVM	81.29
HOG + SVM	84.71
BoVW (SURFC + CLBP) + SVM [25]	85.66
BoVW (SIFT) + SPM + SVM [26]	87.32
Proposed algorithm	**89.73**

four green lines represent those of BoVW using SPM patterns, and the four black lines are the those of BoVW using the proposed DCRP patterns. Overall, the BoVW using SPM pattern is better than tradition BoVW without any space pattern, and the BoVW using DCRP pattern is better than that of SPM pattern. It shows that for commodity images, the introduction of DCRP spatial pattern is beneficial to the improvement of classification accuracy. Besides, as seen from the diagram, the improved BoVW, which adopts MLRE, local feature description of SURF + CVAH, and DCRP spatial patterns, has the highest classification accuracy 89.73%. It shows that the three aspects of improvement to the BoVW model are complementary. Compared with the traditional BoVW and the BoVW adopting SURF + SPM, the classification accuracy of the improved BoVW is increased by 3.60% and 2.33%, respectively.

4.4. Comparison with Other Image Classification Methods. In order to further show the effectiveness of the proposed algorithm (improved BoVW), the classification effect of other algorithms is compared here, as shown in Table 1. LBP and HOG (histograms of oriented gradients) are the method based on the global underlying visual features, and the other three approaches all employ BoVW to describe the intermediate semantic feature. As can be seen from Table 1, the classification effects of the methods using BoVW are better than those based on the global underlying visual features. In [25], the local feature is extracted by SURFC + CLBP, but no spatial model is adopted, which results to loss of spatial information. In [26], the local feature is extracted by SIFT, and the SPM model is also employed. In the proposed method, the local feature is described by SURF (MLRE) + CVAH (MLRE), and the DCRP model is explored to compensate spatial information. Because of good local feature description and spatial pattern, compared with other four methods, the classification accuracy of the proposed method is increased by 8.44%, 5.02%, 4.07%, and 2.41%, respectively.

5. Conclusions

As a subtopic of image processing in the field of e-commerce, commodity image classification should not only solve the common problems in general image classification but also make targeted improvement according to the characteristics of commodity image. To the disadvantage that the traditional BoVW model cannot effectively describe the characteristics of commodity images, this paper proposes an improved BoVW model, which realizes multiscale local region extraction, adopts SURF + CVAH local feature description to add color information, and explores diagonal concentric rectangular pattern to supply spatial information. Experimental results show that the improved BoVW is very suitable for commodity image classification. As providing discriminative features, the improved BoVW proposed in this paper can be also used in image retrieval systems [27]. The subsequent work is to further develop reasonable local feature descriptors and further simplify the model to reduce the computational cost.

Data Availability

The Microsoft PI100 used to support the findings of this study is open dataset, which can be downloaded from: https://pan.baidu.com/s/15nVkXkw06GoFxqs1fVtrVQ using password t29E.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the Heilongjiang Provincial Natural Science Foundation of China (LH2020F008) and Young Innovative Talents Support Project of Harbin University of Commerce (2020CX08).

References

[1] S. Chen and Z. An, "Image retrieval based on image entropy and regional expansion," *International Journal of Control and Automation*, vol. 9, no. 6, pp. 403–410, 2016.

[2] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, p. 1, 2019.

[3] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 23, no. 5, pp. 2019–2032, 2014.

[4] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[5] M. Hidajat, "Annotation based image retrieval using GMM and spatial related object approaches," *International Journal of Control and Automation*, vol. 8, no. 8, pp. 399–408, 2015.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[7] M. A. Stricker and M. Orengo, "Similarity of color images," *Proceedings of SPIE*, vol. 2420, pp. 381–392, 1995.

[8] N. Varish and A. K. Pal, "Content based image retrieval using statistical features of color histogram," in *Proceedings of the 2015 3rd International Conference on Signal Processing, Communications and Networking*, pp. 1–6, Chennai, India, March 2015.

[9] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[10] M. Pietikäinen, T. Ojala, and Z. Xu, "Rotation-invariant texture classification using feature distributions," *Pattern Recognition*, vol. 33, no. 1, pp. 43–52, 2000.

[11] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognition*, vol. 43, no. 3, pp. 706–719, 2010.

[12] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[13] H. Lategahn, S. Gross, T. Stehle, and T. Aach, "Texture classification by modeling joint distributions of local patterns with Gaussian mixtures," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1548–1557, 2010.

[14] S. Fekri-Ershad, "Developing a gender classification approach in human face images using modified local binary patterns and tani-moto based nearest neighbor algorithm," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 12, no. 4, pp. 1–12, 2019.

[15] P. Shanmugavadivu, P. Sumathy, and A. Vadivel, "FOSIR: fuzzy-object-shape for image retrieval applications," *Neurocomputing*, vol. 171, pp. 719–735, 2016.

[16] H. Wu and S. Yan, "Computing invariants of Tchebichef moments for shape based image retrieval," *Neurocomputing*, vol. 215, pp. 110–117, 2016.

[17] E. Sokic and S. Konjicija, "Phase preserving fourier descriptor for shape-based image retrieval," *Signal Processing: Image Communication*, vol. 40, pp. 82–96, 2016.

[18] Wikipedia. Bag-of-words model in computer vision [EB/OL]. http://en.wikipedia.org/wiki/Bagof-words_model_in_coniputer_vision.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Convolutional neural networks," *Advances in Neural Imagenet Classification with Deep Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[20] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[22] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[23] Q. Meng and Z. Lv, "An improved SIFT algorithm for image registration based realization of the vision figure," *International Journal of Control and Automation*, vol. 9, no. 6, pp. 51–58, 2016.

[24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[25] Z. Zhang and Z. Ju, "Multi-feature fusion fruit and vegetable image classification based on bag of feature model," *Electronic Science and Technology*, vol. 31, no. 1, pp. 1–6, 2019.

[26] H. Zhang, S. Liu, B. Zhang, and J. Wang, "Natural scene recognition based on spatial pyramid integrated bag-of-visual-words model," *Journal of Shanghai Jiaotong University*, vol. 50, no. 6, pp. 902–909, 2016.

[27] N. T. Bani and S. Fekri-Ershad, "Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains," *The Electronic Library*, vol. 37, no. 4, pp. 650–666, 2019.

Hindawi
Complexity
Volume 2021, Article ID 5570575, 10 pages
https://doi.org/10.1155/2021/5570575

Research Article

An Improved Integrated Scheduling Algorithm with Process Sequence Time-Selective Strategy

Zhen Wang, Xiaohuan Zhang, and Gang Peng

School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China

Correspondence should be addressed to Xiaohuan Zhang; zhangxiaohuan@hzu.edu.cn

Received 16 February 2021; Revised 25 February 2021; Accepted 27 February 2021; Published 9 March 2021

Academic Editor: Abd E.I.-Baset Hassanien

The integrated scheduling algorithm of process sequence time-selective strategy (ISAOPSTSS) is an advanced algorithm in the field of integrated scheduling. The proposed algorithm points out the shortcomings of the process sequence time-selective strategy. Generally, there are too many "trial scheduling" times. The authors propose that there is no need to make "trial scheduling" at every "quasi-scheduling time point." In fact, the process scheduling scheme can be obtained by trial scheduling on some "quasi-scheduling time points." The scheduling result is the same as that of the sequence timing strategy. The proposed algorithm reduces the runtime of scheduling algorithm and improves the performance of the algorithm without reducing the optimization effect.

1. Introduction

Scheduling is a key factor affecting the production efficiency of manufacturing industry. Effective scheduling optimization algorithm can maximize the production efficiency on the premise of satisfying the constraints of enterprise orders, equipment, and other hardware and software resources. At present, scholars have done a lot of research on the workshop scheduling problem, and most scheduling problems are mainly divided into job shop [1–10] and flow shop [11–19]. These algorithms are mainly for scheduling optimization when the workpiece is first processed and then assembled into the product. At present, consumers have more and more demand for individual products, and manufacturing factories will face more and more orders for multivariety and small batch products. This kind of situation if the production is still in processing after assembling according to the traditional mode of production would split the product internal parallel processing and assembly relations and reduce the production efficiency. In order to seek solutions for this new research field, Zhi-Qiang et al. [20] proposed an integrated scheduling algorithm that simultaneously promotes product processing and assembly, developed a series of scheduling optimization algorithms, and expanded many new research fields.

2. Review of Related Studies

At present, in the field of general integrated scheduling research, the following research studies have been mainly carried out.

Zhi-Qiang et al. [20] firstly pointed out the important position of critical path in the process tree and emphasized that the scheduling of processes with vertical relationship in the process tree is closely related to the final scheduling result. Yang et al. [21] put forward the strategy of layer priority, short time, long path, and dynamic adjustment and pointed out that adding the parallel relationship between processes with horizontal relationship in the process tree can make the scheduling result better. Yang et al. [22] pointed out that reference [20] paid attention to the vertical structure of product tree structure and ignored the horizontal parallel processing of the same equipment process; at the same time, it is pointed out that reference [21] focuses on the horizontal structure of the product tree structure and considers the vertical path on the basis of the horizontal layer. However, the strategy of emphasizing horizontal and neglecting vertical does not conform to the mechanism of vertical-oriented product scheduling. A scheduling scheme with vertical as the main factor and horizontal as the main factor is proposed. The advantages of the algorithm are as follows: on the basis

of both vertical and horizontal, the vertical scheduling is further optimized, which is in line with the idea that the integrated scheduling is mainly vertical. The disadvantages are as follows: although the dynamic critical path idea is used to solve the problem that the serial process and parallel process are pushed forward at the same time, the idea of the algorithm is too macro, and due to the restriction of some factors, it is impossible to consider the tightness between serial processes.

Xin et al. [23] determined the scheduling sequence of process according to the length of path according to the scheduling algorithm in document [20], and forms parallel 4 processing among groups, resulting in more idle time. An integrated scheduling algorithm based on device idle event driven is proposed. On the basis of reference [23], Xin et al. [24] proposed to further optimize the scheduling results by using rollback strategy to schedule the processes with long path of parent node first; the advantages are as follows: it increases the utilization rate of equipment, maximizes the "equipment busy" principle, reduces the idle time of equipment, and makes the process more compact. The disadvantage is that "device-driven events" always look for processes in the current schedulable parallel processes. From the perspective of processes, the algorithm can be regarded as an improved "layer first" scheduling under the "leaf alignment" mode. This algorithm increases the parallelism and the processing waiting time between the serial processes and ignores the impact of vertical scheduling optimization on the scheduling results.

To sum up, the current research can optimize the parallel scheduling of processes in products, but the scheduling optimization of serial processes needs to be improved while considering the parallelism. On the basis of the above research, Xie et al. [25] proposed a time-selective integrated scheduling algorithm considering the compactness of serial operations (ISAOPSTSS). The algorithm not only inherits the advantages of the current algorithm to ensure the parallelism between processes but also optimizes the compactness between serial processes on the basis of it and further emphasizes the scheduling idea of vertical optimization. It avoids the disadvantages of previous algorithms and optimizes the scheduling results. However, ISAOPSTSS in determining the scheduling scheme of the process is more complex, resulting in operation redundancy. This paper proposes an improved algorithm, which can reduce execution time of algorithm and improve the performance of the algorithm without reducing the optimization effect of the algorithm.

3. Problem Description and Analysis

The integrated scheduling problem is to study how to schedule the processes to minimize the product completion time when the product is in the production mode of assembly while processing. Among them, the researchers regard the processing and assembly of each process as a whole, collectively referred to as processing. The processing time, processing equipment, and partial order of each process in the product are clearly indicated by the product processing tree. The integrated dispatching shall meet the following requirements:

(1) Each process can only be processed on one machine

(2) Each time a machine can only process one process

(3) If and only if all the preprocesses of a process are in the state of finished processing (or no preprocesses), the process can be processed

(4) The processing of a certain process cannot be interrupted

(5) The difference between the processing end time of the latest finishing process and the processing start time of the earliest starting process is the total processing time of the product

4. Analysis and Design of Scheduling Strategy

4.1. Analysis of Improved Process Sequence Time-Selective Strategy. As shown in Figure 1, the reverse order process tree of product A proposed in ISAOPSTSS is analyzed as follows:

Step 1: apply the sequencing strategy of process sequence to sort the processes in product A. According to the product process tree as shown in Figure 2, first calculate the path length of all the leaf node processes, and the results were as follows: A10: 10, A9: 16, A5: 21, A8: 20, and A11: 9. Therefore, all nodes on the path where A5 is located are selected as the first process sequence, and these processes are added to the process queue Q_u. At this time, the sequence of processes in queue Q_u is A1, A2, A3, A4, and A5; at the same time, delete these processes in the process tree of product A. At this time, the processing tree of product A becomes a forest composed of multiple subtrees. Next, the path length of leaf nodes in these subtrees is calculated in turn, and the results were as follows: A10: 9, A9: 1, A8: 20, and A11: 8. Select all nodes on the path where A8 is located as the first process sequence, and add these processes to the process queue Q_u, and the sequence of processes in queue Q_u is A1, A2, A3, A4, A5, A6, A7, and A8, and these processes are deleted in the process tree of product A. By analogy, the sequence of processes in the process queue Q_u corresponding to the process tree of final product A is A1, A2, A3, A4, A5, A6, A7, A8, A10, A11, and A9, and this sequence will be used as the scheduling sequence of processes.

Step 2: schedule the longest operation sequence in the queue Q_u to the previous operation to form the initial scheduling scheme, as shown in Figure 3.

Step 3: the improved sequencing strategy of process sequence is used to schedule the remaining processes in the process queue Q_u in turn. First, the process A6 is scheduled. The earliest start processing time of A6 is 1, and the processing device is $M3$. Because A2 and A5 have been scheduled on $M3$ device. The "quasi-scheduling time points" of A6 are the earliest start processing time 1, the end processing time 4 of process A2, and the end

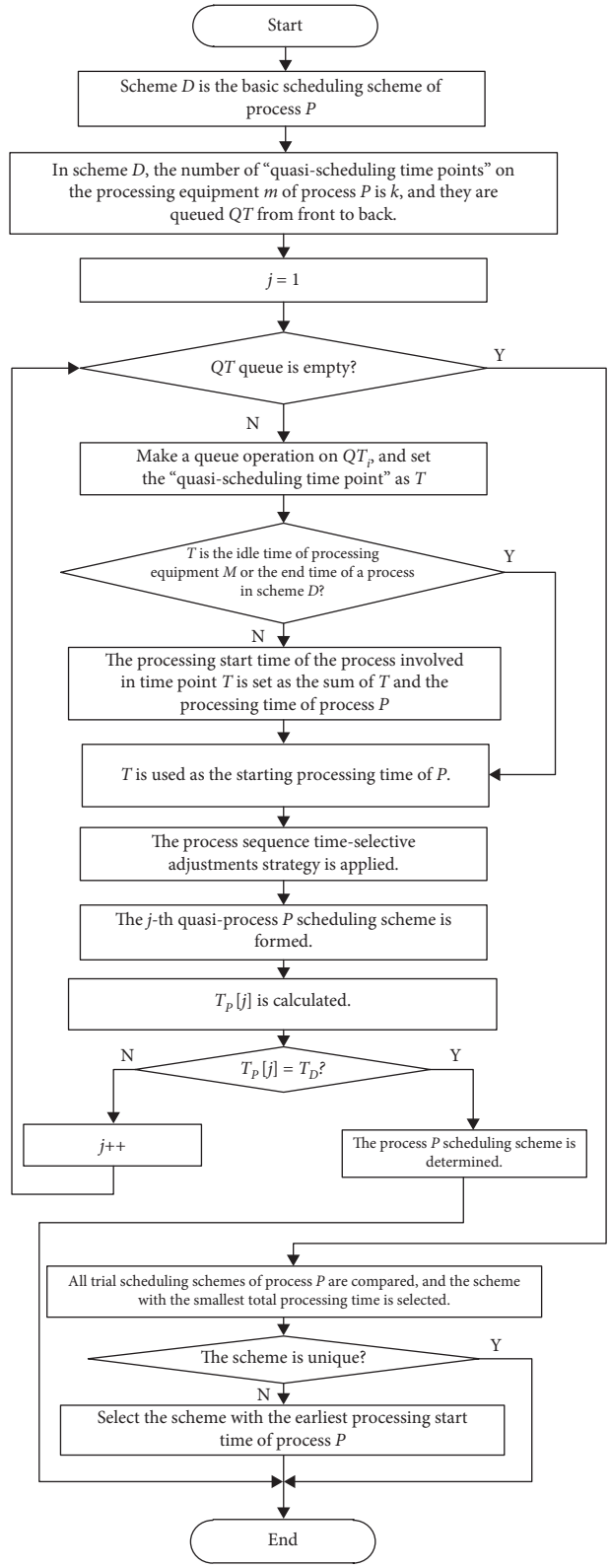


Figure 1: The flowchart of the process sequence time-selective scheduling strategy.

processing time 21 of process A5. Try to schedule A6 at these three time points, and get three trial scheduling schemes as shown in Figure 4.

In the A6 "quasi-scheduling scheme" shown in Figure 4, ISAOPSTSS calculated the total processing time of each "quasi-process scheduling scheme" formed at each "quasi-scheduling time point" and the scheme shown in Figure 4(b)

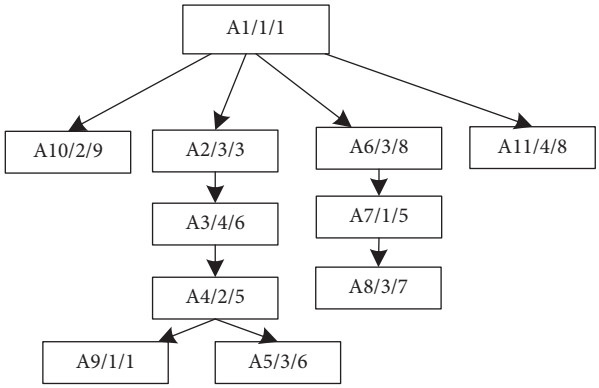


Figure 2: Process tree of product A.

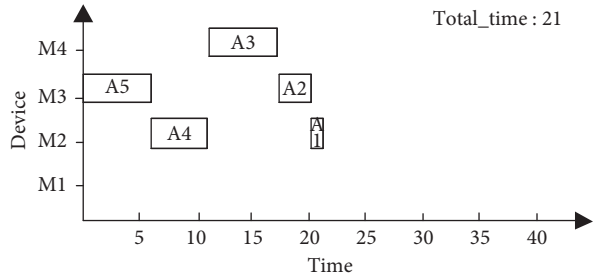


Figure 3: Gantt chart of initial scheduling scheme for product A.

is selected as the scheduling scheme of A6. scheme with the least total processing time as the A6 scheduling scheme.

In fact, according to the characteristics of the process sequence time-selective algorithm, it is not necessary to calculate the total processing time of each "quasi-process scheduling scheme" formed on each "quasi-scheduling time point."

First, the total processing time of the current basic scheduling scheme is T_D, and the processing time of the current scheduling process is t. The total processing time of the current process P scheduling is T_p.

$$T_D \leq T_p \leq T_D + t. \tag{1}$$

According to formula (1), the total processing time of the current process P quasi-scheduling scheme is T_p, the minimum processing time is T_D, and the maximum processing time is the sum of the total processing time t of the current basic scheduling scheme and the processing time t of the current scheduling process. It can be seen that if the "quasi-scheduling time point" is used to schedule the processes in the order from front to back and if the total processing time of a quasi-process scheduling scheme is T_D, it can be determined that the quasi-process scheduling scheme on the quasi-scheduling time point is the current process scheduling scheme, and there is no need to calculate the total processing time of the quasi-process scheduling scheme on the quasi-scheduling time point after the time point; of course, the total processing time T_D of the "quasi-operation scheduling scheme" generated from the "quasi-scheduling time point" is equal to the total processing time T_D of the basic scheduling scheme; however, due to the fact that the "quasi-scheduling time point" is scheduled in the order from the front to the back in the trial scheduling process, the former "quasi-scheduling time point" is better than the latter in the case of the same total processing time from the perspective of interprocess compactness.

Therefore, the current process scheduling scheme will be discussed in the following two situations:

(1) If the total processing time of the "quasi-process scheduling scheme" generated at the current "quasi-scheduling time point" is greater than that of the current basic scheduling scheme, the total processing time of the "quasi-process scheduling scheme" generated at the next "quasi-scheduling time point" will continue to be determined.

(2) If the total processing time of the "quasi-process scheduling scheme" generated from the current "quasi-scheduling time point" is equal to the total processing time of the current basic scheduling scheme, the calculation of the "quasi-scheduling time point" behind will be stopped, and the "quasi-process scheduling scheme" generated from the current "quasi-scheduling time point" will be taken as the current process scheduling scheme.

4.2. Algorithm Design of Improved Process Sequence Time-Selective Scheduling Strategy. The specific steps of the improved algorithm are as follows:

Step 1: set the basic scheduling scheme of process P as D.

Step 2: in scheme D, k "quasi-scheduling time points" on the processing equipment m of process P are found, and they are queued QT from front to back, $j = 1$.

Step 3: judge whether the QT queue is empty. If it is not empty, make a queue operation on the QTi queue. Take out the "quasi-scheduling time point" T and go to step 4. If it is empty, go to step 12.

Step 4: judge whether T is the idle time of processing equipment M or the end time of a process in scheme D. If not, go to step 5 and if so, go to step 6.

Step 5: the processing start time of the process involved in time point T (the process being processed at time point T) is set as the sum of T and the processing time of process P.

Step 6: time point T is used as the starting processing time of process P to schedule process P.

Step 7: the process sequence time-selective adjustment strategy [25] is used to adjust the process affected by scheduling process P, forming the j-th quasi-process P scheduling scheme.

Step 8: the total processing time $T_P[j]$ of the j-th quasi-process P scheduling scheme is calculated.

Step 9: judge whether $T_P[j] = T_D$ is true. If yes, go to step 10; otherwise, go to step 11.

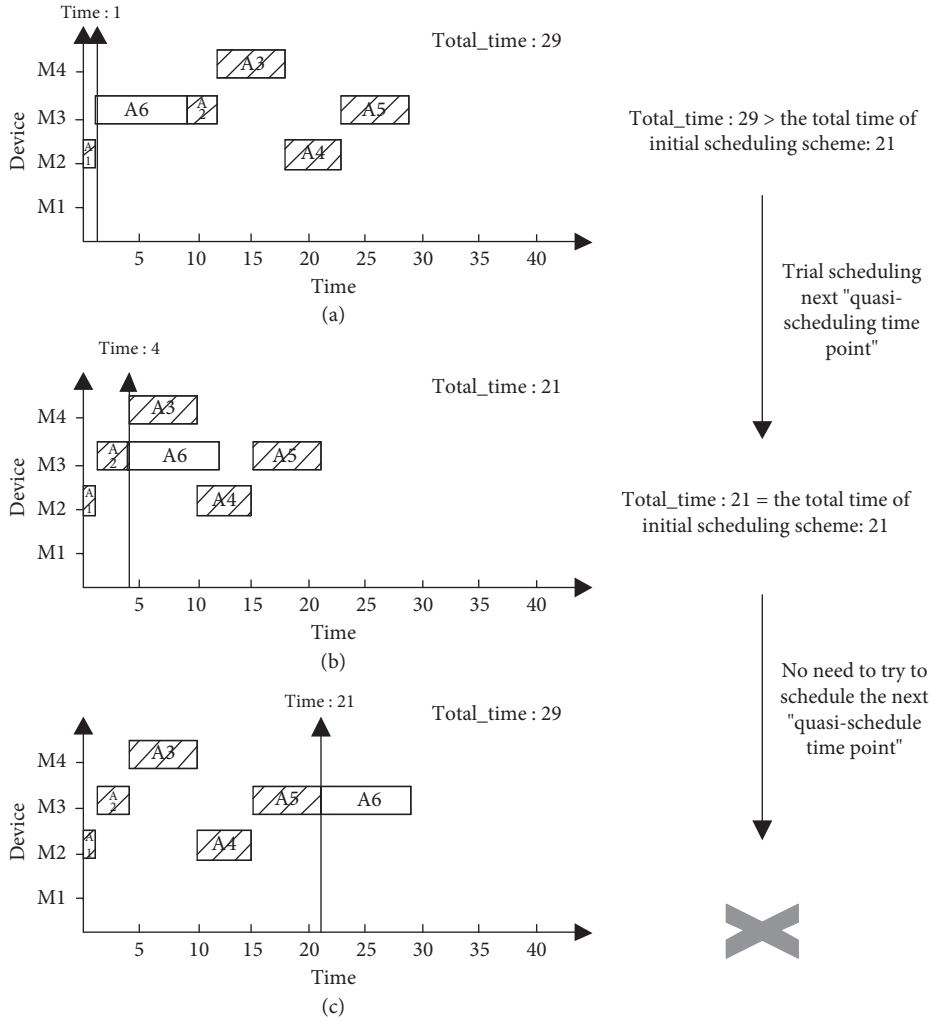


Figure 4: Gantt chart of each trial scheduling scheme of process A6. (a) The scheduling scheme is obtained by trial scheduling at quasi-scheduling time point "1." (b) The scheduling scheme is obtained by trial scheduling at quasi-scheduling time point "4." (c) The scheduling scheme is obtained by trial scheduling at quasi-scheduling time point "21."

Step 10: the scheduling scheme obtained at time point T is taken as the scheduling scheme of process P, and go to step 14.

Step 11: j++, go to Step 3.

Step 12: the total processing time of j quasi-process P scheduling schemes is compared, and the scheme with the smallest total processing time is selected.

Step 13: judge whether the scheme is unique. If it is unique, select it. If it is not unique, select the scheme with the earliest processing start time of process P.

Step 14: exit.

The flowchart of the process sequence time-selective scheduling strategy is shown in Figure 1.

5. Algorithm Design

The implementation steps of the improved algorithm are as follows:

Step 1: the reverse order processing tree is obtained by reversing the partial order relationship of processing processes in the processing tree.

Step 2: the operation queue Q_u is obtained by using the process sequence sorting strategy.

Step 3: all processes in the longest process sequence on Q_u are queued and scheduled to form the initial scheduling scheme P_0.

Step 4: $i = 1$.

Step 5: judge whether Q_u is empty. If it is empty, go to step 8; otherwise, go to step 6.

Step 6: the queue Q_u is queued to obtain the current scheduling process P; the processing time is t, and the processing equipment is M.

Step 7: the improved process sequence time-selective strategy is applied to schedule P, and the process P scheduling scheme is obtained; i++, go to step 5.

Step 8: form product scheduling Gantt chart and output.

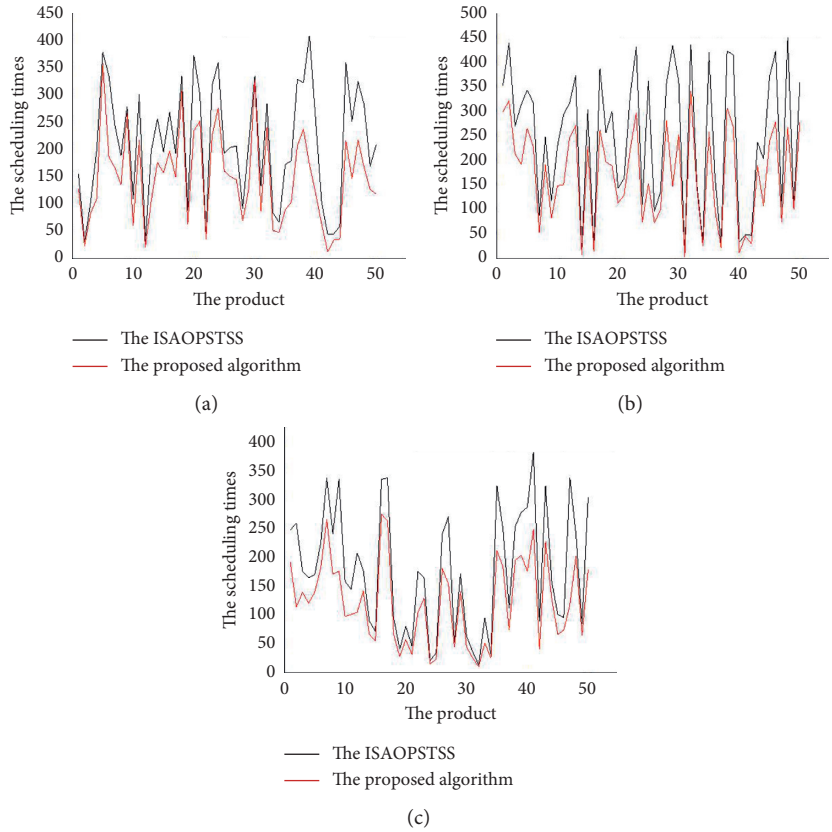


Figure 5: Comparison of scheduling times when the total number of processes is 30. (a) The scheduling times of 30 processes in 3 devices. (b) The scheduling times of 30 processes in 6 devices. (c) The scheduling times of 30 processes in 9 devices.

6. Experimental Method

In order to verify the performance of the algorithm, a group of experiments is designed for comparison. The algorithm performance is affected by the structure of product process tree, the number of processing equipment, and the processing time. Therefore, the proposed algorithm is compared with the ISAOPSTSS algorithm from the perspective of different scale parameters. First, each experiment will randomly generate 50 products. The parameters in the process tree are randomly generated. The parameters are as follows: the structure of the process tree (including the total number of layers of the process tree, the number of processes in each layer, and the relationship between the front and back of the process), the processing time and equipment number of the process in the process tree, and the total number of processing equipment of the process. Each group of experiments will randomly generate 50 products because the product structure is random, which can effectively prove the effectiveness of the algorithm in different cases. The above two algorithms are implemented in C++ language by dev c++ 4.9.9.2. Schedule the randomly generated product process trees, set counters in the two algorithms to monitor the times of "trial scheduling" in each algorithm, and record them for comparison. Since the algorithm proposed in this paper is an improvement on the algorithm proposed in ISAOPSTSS, if the number of "trial scheduling" of the algorithm proposed in this paper is less than that in ISAOPSTSS under different parameter conditions, the effectiveness of the algorithm proposed in this paper can be proved.

7. Results and Discussion

Six groups of experiments were designed as follows. In Experiment 1, as shown in Figure 5, the proposed algorithm is compared with the algorithm in ISAOPSTSS. In order to compare, 50 product process trees were randomly generated, the total number of processes is 30, and the total number of processing equipment is 3, 6, and 9, respectively. In Experiment 2, as shown in Figure 6, the proposed algorithm is compared with the algorithm in ISAOPSTSS. The comparison data are used to randomly generate 50 groups, the total number of processes is 50, and the total number of processing equipment is 3, 6, and 9, respectively. In Experiment 3, as shown in Figure 7, the proposed algorithm is compared with the algorithm in ISAOPSTSS. The comparison data are used to randomly generate 50 groups, the total number of processes is 80, and the total number of processing equipment is 3, 6, and 9, respectively. In Experiment 4, as shown in Figure 8, the proposed algorithm is compared with the algorithm in ISAOPSTSS. The comparison data are used to randomly generate 50 groups, the total number of processes is 100, and the total number of processing equipment is 3, 6, and 9, respectively. As shown in Figure 9, Experiment 5 shows the average scheduling

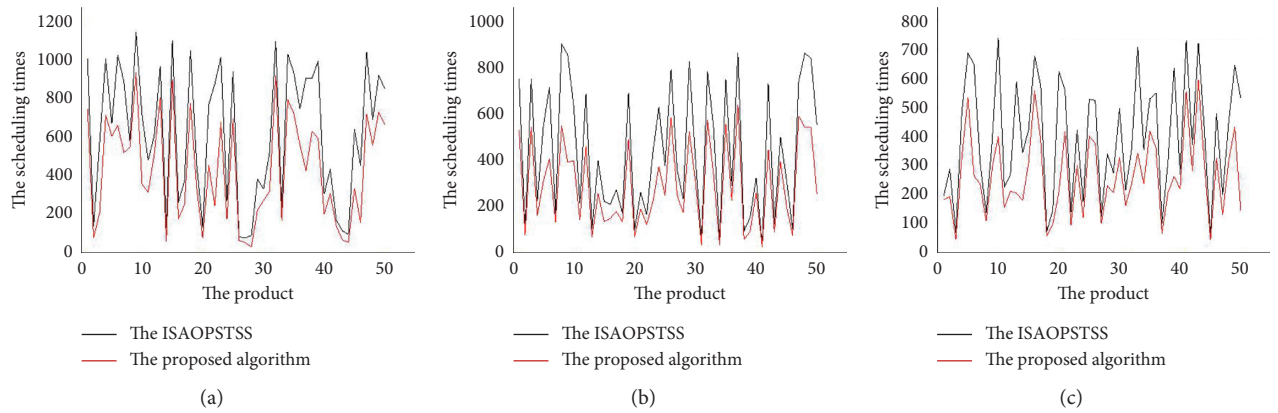


Figure 6: Comparison of scheduling times when the total number of processes is 50. (a) The scheduling times of 50 processes in 3 devices. (b) The scheduling times of 50 processes in 6 devices. (c) The scheduling times of 50 processes in 9 devices.

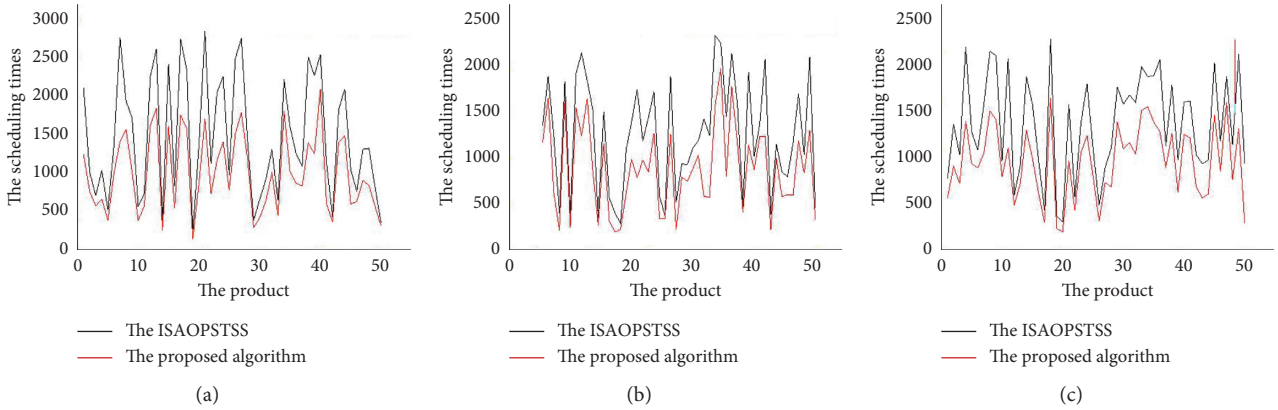


Figure 7: Comparison of scheduling times when the total number of processes is 80. (a) The scheduling times of 80 processes in 3 devices. (b) The scheduling times of 80 processes in 6 devices. (c) The scheduling times of 80 processes in 9 devices.

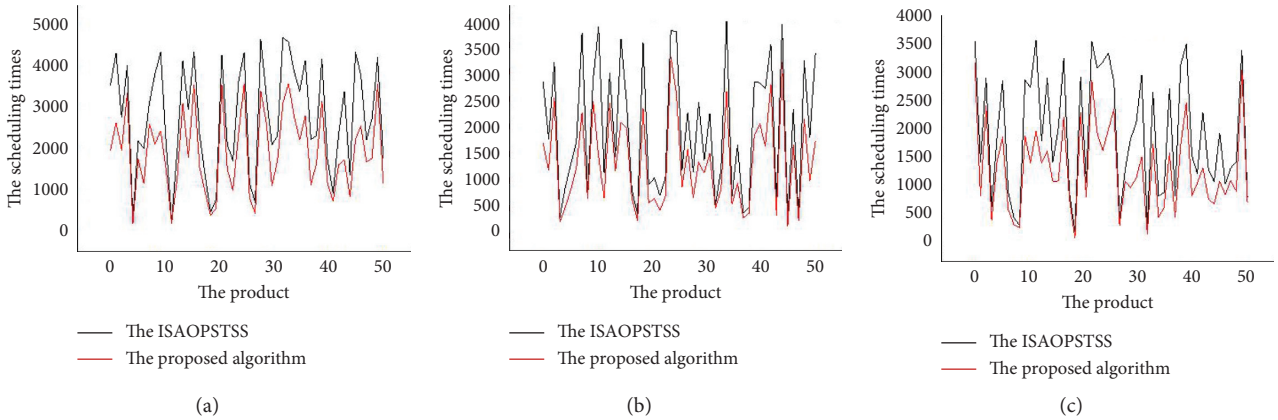


Figure 8: Comparison of scheduling times when the total number of processes is 100. (a) The scheduling times of 100 processes in 3 devices. (b) The scheduling times of 100 processes in 6 devices. (c) The scheduling times of 100 processes in 9 devices.

times comparison between the proposed algorithm and ISAOPSTSS algorithm when the total number of processes is 30, 50, 80, and 100, respectively. As shown in Figure 10, Experiment 6 shows that in Experiment 1, when the total number of devices is 3, 6, and 9, respectively, the average scheduling times of the proposed algorithm are compared with those of ISAOPSTSS.

Analysis of the above experimental data shows that the number of trial scheduling times of the algorithm in this paper is significantly reduced compared with the number of

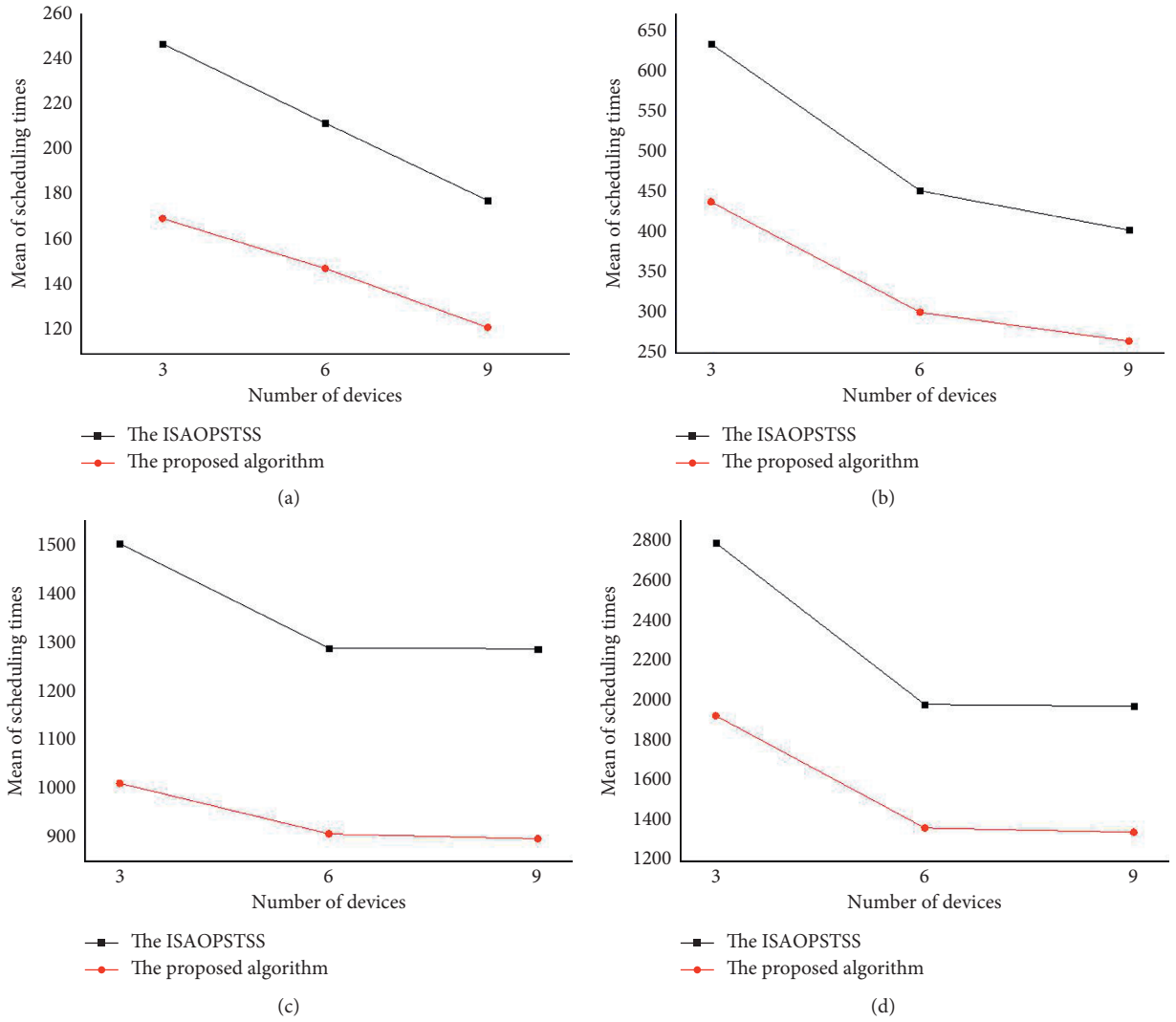


Figure 9: The comparison chart of the average scheduling times when the total number of processes is 30, 50, 80, and 100, respectively. (a) Comparison chart of mean scheduling times when the number of processes is 30. (b) Comparison chart of mean scheduling times when the number of processes is 50. (c) Comparison chart of mean scheduling times when the number of processes is 80. (d) Comparison chart of mean scheduling times when the number of processes is 100.

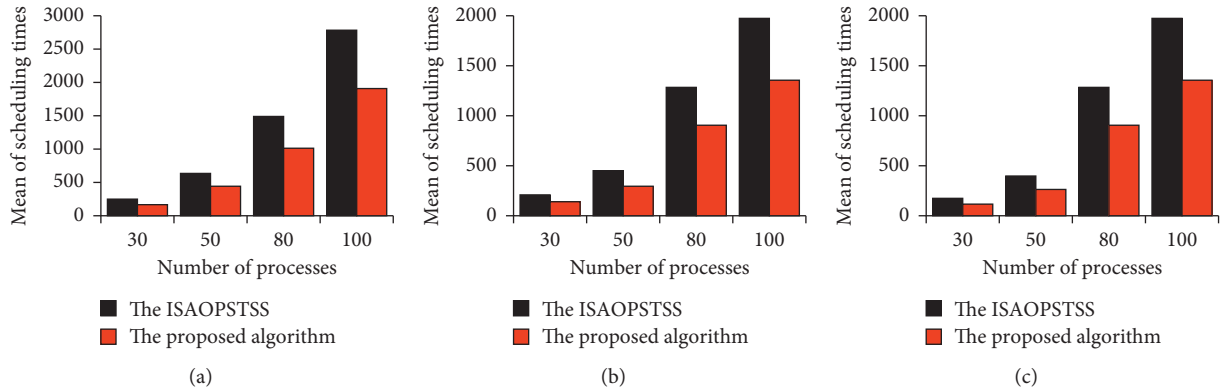


Figure 10: The comparison chart of the average scheduling times when the total number of devices is 3, 6, and 9, respectively. (a) Comparison chart of mean scheduling times when the number of devices is 3. (b) Comparison chart of mean scheduling times when the number of devices is 6. (c) Comparison chart of mean scheduling times when the number of devices is 9.

pilot scheduling in ISAOPSTSS, and the reduction ratio is about 30%.

8. Conclusions

The algorithm proposed in this paper is a further optimization of the operation sequence timing algorithm. On the premise of ensuring the optimization results, it simplifies the algorithm steps, reduces the algorithm execution time, and improves the algorithm performance. At present, the ISAOPSTSS algorithm has been applied in batch processing scheduling and two-job-shop scheduling and other fields. It may be the next step to apply the proposed algorithm in these fields to improve the performance of the algorithm.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the Project of Educational Commission of Guangdong (grant no. 2019KTSCX177), the Science and Technology Planning Project of Guangdong (grant no. 2020A1414010235), the Science and Technology Planning Project of Huizhou (grant no. 2020SC0306023), and the Professorial and Doctoral Scientific Research Foundation of Huizhou University (grant nos. 2019JB014 and 2018JB007).

References

[1] Z. Cao, L. Zhou, B. Hu, and C. Lin, "An adaptive scheduling algorithm for dynamic jobs for dealing with the flexible job shop scheduling problem," *Business & Information Systems Engineering*, vol. 61, no. 3, pp. 299–309, 2019.

[2] P. Pongchairerks, "A two-level metaheuristic algorithm for the job-shop scheduling problem," *Complexity*, vol. 2019, Article ID 8683472, 11 pages, 2019.

[3] A. Jamili, "Job shop scheduling with consideration of floating breaking times under uncertainty," *Engineering Applications of Artificial Intelligence*, vol. 78, pp. 28–36, 2019.

[4] J. Lin, "Backtracking search based hyper-heuristic for the flexible job-shop scheduling problem with fuzzy processing time," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 186–196, 2019.

[5] Z. C. Li, B. Qian, R. Hu, L. L. Chang, and J. B. Yang, "An elitist nondominated sorting hybrid algorithm for multi-objective flexible job-shop scheduling problem with sequence-dependent setups," *Knowledge-Based Systems*, vol. 173, pp. 83–112, 2019.

[6] K. Gao, F. Yang, M. Zhou, Q. Pan, and P. N. Suganthan, "Flexible job-shop rescheduling for new job insertion by using discrete jaya algorithm," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1944–1955, 2019.

[7] L. Sun, L. Lin, M. Gen et al., "A hybrid cooperative coevolution algorithm for fuzzy flexible job shop scheduling," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, p. 1, 2019.

[8] B. Wang, H. Xie, X. Xia et al., "A NSGA-II algorithm hybridizing local simulated-annealing operators for a Bi-criteria robust job-shop scheduling problem under scenarios," *IEEE Transactions on Fuzzy Systems*, vol. 27, 2018.

[9] M. M. Ahmadian, A. Salehipour, and T. C. E. Cheng, "A meta-heuristic to solve the just-in-time job-shop scheduling problem," *European Journal of Operational Research*, vol. 288, 2020.

[10] M. Abedi, R. Chiong, N. Noman et al., "A multi-population, multi-objective memetic algorithm for energy-efficient job-shop scheduling with deteriorating machines," *Expert Systems with Applications*, vol. 157, Article ID 113348, 2020.

[11] M. Qin, R. Wang, Z. Shi et al., "A genetic programming-based scheduling approach for hybrid flow shop with a batch processor and waiting time constraint," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 99, pp. 1–12, 2019.

[12] D. Ferone, S. Hatami, E. M. GonzálezkMeira et al., "A biased-randomized iterated local search for the distributed assembly permutation flow hop problem," *International Transactions in Operational Research*, vol. 27, no. 3, 2020.

[13] S. Aqil and K. Allali, "Local search metaheuristic for solving hybrid flow shop problem in slabs and beams manufacturing," *Expert Systems with Applications*, vol. 162, Article ID 113716, 2020.

[14] P. Valledor, A. Gomez, P. Priore et al., "Modelling and solving rescheduling problems in dynamic permutation flow shop environments," *Complexity*, vol. 2020, Article ID 2862186, 17 pages, 2020.

[15] M. Fazayeli, M. R. Aleagha, R. Bashirzadeh et al., "A hybrid meta-heuristic algorithm for flowshop robust scheduling under machine breakdown uncertainty," *International Journal of Computer Integrated Manufacturing*, vol. 29, pp. 1–11, 2016.

[16] A. Hasani and S. M. H. Hosseini, "A bi-objective flexible flow shop scheduling problem with machine-dependent processing stages: trade-off between production costs and energy consumption," *Applied Mathematics and Computation*, vol. 2020, p. 386, Article ID 125533, 2020.

[17] K. Geng, L. Liu, C. Ye et al., "Bi-objective re-entrant hybrid flow shop scheduling considering energy consumption cost under time-of-use electricity tariffs," *Complexity*, vol. 2020, Article ID 8565921, 17 pages, 2020.

[18] A. Costa, F. V. Cappadonna, and S. Fichera, "Minimizing makespan in a flow shop sequence dependent group scheduling problem with blocking constraint," *Engineering Applications of Artificial Intelligence*, vol. 89, Article ID 103413, 2020.

[19] M. S. Nagano, J. V. S. Robazzi, and C. P. Tomazella, "An improved lower bound for the blocking permutation flow shop with total completion time criterion," *Computers & Industrial Engineering*, vol. 146, Article ID 106511, 2020.

[20] X. Zhi-Qiang, L. Sheng-Hui, and P.-L. Qiao, "Dynamic job-shop scheduling algorithm based on ACPM and BFSM," *Journal of Computer Research and Development*, vol. 40, no. 7, pp. 977–983, 2003.

[21] Z. Xie, J. Yang, G. Yang et al., "Dynamic job-shop scheduling algorithm with dynamic set of operation having priority," *Chinese Journal of Computers*, vol. 31, no. 3, pp. 502–508, 2008.

[22] Z.-Q. Xie, J. Yang, Y. Zhou, D.-L. Zhang, and G.-Y. Tan, "Dynamic critical paths multi-product manufacturing scheduling algorithm based on operation set," *Chinese Journal of Computers*, vol. 34, no. 2, pp. 406–412, 2011.

[23] Z. Xie, Y. Xin, and J. Yang, "Integrated scheduling algorithm based on event-driven by machines' idle," *Journal of Mechanical Engineering*, vol. 47, no. 11, pp. 139–147, 2011.

[24] Z. Xie, Y. Xin, and J. Yang, "Machine-driven integrated scheduling algorithm with rollback- preemptive," *Automatica Sinica*, vol. 37, no. 11, pp. 1332–1343, 2011.

[25] Z. Xie, X. Zhang, Y. Gao, and Y. Xin, "Time-selective integrated scheduling algorithm considering the compactness of serial processes," *Journal of Mechanical Engineering*, vol. 54, no. 6, pp. 191–202, 2018.

Hindawi
Complexity
Volume 2021, Article ID 6691571, 16 pages
https://doi.org/10.1155/2021/6691571

Research Article

Intelligent and Smart Irrigation System Using Edge Computing and IoT

M. Safdar Munir, Imran Sarwar Bajwa, Amna Ashraf, Waheed Anwar, and Rubina Rashid

Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

Correspondence should be addressed to Imran Sarwar Bajwa; imran.sarwar@iub.edu.pk

Received 17 December 2020; Revised 8 February 2021; Accepted 18 February 2021; Published 28 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

Smart parsimonious and economical ways of irrigation have build up to fulfill the sweet water requirements for the habitants of this world. In other words, water consumption should be frugal enough to save restricted sweet water resources. The major portion of water was wasted due to incompetent ways of irrigation. We utilized a smart approach professionally capable of using ontology to make 50% of the decision, and the other 50% of the decision relies on the sensor data values. The decision from the ontology and the sensor values collectively become the source of the final decision which is the result of a machine learning algorithm (KNN). Moreover, an edge server is introduced between the main IoT server and the GSM module. This method will not only avoid the overburden of the IoT server for data processing but also reduce the latency rate. This approach connects Internet of Things with a network of sensors to resourcefully trace all the data, analyze the data at the edge server, transfer only some particular data to the main IoT server to predict the watering requirements for a field of crops, and display the result by using an android application edge.

1. Introduction

Agriculture is the major resource of living wage in Pakistan. A smart, intelligent, and fully automated agricultural system was required and extremely desirable in some last decades when our population grew exponentially in comparison to the natural resources we have in our country, Pakistan. For this purpose, an IoT-based smart watering system has been achieved in the recent years of constant threat of losing water. This agricultural industry has two particulars. The plastic tunnel farming is divided into low, high, and walk-in tunnels. It is convenient to sow, spray, and harvest in the high tunnel than in low and walk-in tunnels due to its broader size. Traditional farming, on the contrary, is the most unpredictable and becomes the cause of more water wastage. The issue we are going to deal with in this paper regarding smart irrigation is any application designed and used for the smart watering system still needs to be more efficient and timely. Technically, it means that just cloud computing is not enough for a large-scale IoT application. There should be something like more efficient and fast application using a better architecture to handle different types of data coming from different sources (sensors). The main purpose of a quick and smart irrigation system is the consumption of water so frugally to execute the need of water more timely for a field of plants and to save inadequate sweet water reservoirs. To handle this rigorous matter, many sensor-based smart irrigation systems with their mobile applications have been designed in different times, but still, there is a question on their reliability when data grow and thus the latency rate of IoT devices. Like in preceding papers, the input parameters humidity, temperature, soil moisture, and light intensity were used, and a decision of watering plants or not was made on the basis of a fuzzy logic [1]. The same fuzzy logic has been applied to many healthcare systems, in which use of biosensors helped monitoring temperature, blood pressure, oxygen, and infection status of the wound [2]. Similarly, in fire alarming applications, this

technology helped a lot in 2018 [3] and 2019 [4]. Now, we come up with a new technology that is the combination of machine learning technique and semantics for some input parameters such as climate type, crop type, and soil type with the sensors' output: temperature, humidity, and soil moisture.

A smart irrigation system with the application of edge computing is required because the research studies on irrigation systems until now are not much efficient that they could not be implemented on large-scale systems and have less efficiency due to overburdenized sensors for all sensing data. So, a new intelligent and smart system should be designed.

Our research found some grounds due to which improvements in the existing system are mandatory:

(i) Existing smart irrigation systems either spotlighting on lesser parameters like soil moisture, air moisture/humidity or they are presenting a fuzzy logic (implemented in matlab) to produce an output decision or some are using simple machine learning algorithm to predict about water need for plants. A system which does not encounter the latency rate cannot provide the reliable solution.

(ii) Skipping important parameters such as soil strata and crop type can lead to an imperfect watering system for plants.

(iii) Unwanted data loading on the IoT server due to continuous throw of sensor data becomes a cause of less efficiency of the IoT server. An intelligent irrigation system should never halt due to overburden of data.

(iv) As newest expertise has come into sight due to progression in each and every field, therefore, we also have to change our classical method of irrigation to advanced, smart, and perfect and simple knowledge database for plant's data to powerful ontology-based semantics.

There are some main aspects, which we are going to concentrate on in our anticipated approach:

(i) Three sensors are used in our approach: a soil moisture sensor, humidity and temperature sensor, and light sensor. Furthermore, ontology is used for plant species data, different soil types, and different climate types.

(ii) This approach focuses on an intelligent technique, i.e., machine learning, to decide watering requirements for a particular plant, and by considering many other suitable parameters for the plant growth, i.e., climate, weather, and soil type, we are going to design a smart irrigation system in a different and more efficient way.

(iii) Our proposed smart system by design focuses on system reliability as if a sensor for some reason is not working at a particular time and was working an hour before, then the value it measured before an hour will be used by our trained model to produce the result because no drastic change can occur in other parameters in just an hour. It makes our system user friendly and more efficient.

(iv) The proposed approach is structured to come upon the problems of the obsolete irrigation method smartly.

2. Related Works

Traditional tunnel farms, all over the world, use drip irrigation or a sprinkler irrigation method. These are better than normal flooding methods. Various irrigation methods provide different water consumption levels and energy competence [5]. The surface irrigation and level irrigation methods provide low water and energy efficiency. The subirrigation, overhead irrigation, and sprinkler irrigation methods provide low-to-medium efficiency. The sprinkler and drip irrigation methods provide similar energy efficiency, but drip irrigation is more water efficient than sprinkler irrigation [6].

To increase crop production and decrease costs efficiently, the management of freshwater smartly is indispensable. The powerful use of technologies provides the precise amount of water required for plants. The SWAMP project [7] in Europe has developed an IoT-based smart water management platform for ideal irrigation with a proactive approach on four pilots in Brazil and Europe. The SWAMP architecture, the platform, and the system deployed presented by the European people include a performance analysis of FIWARE components. They aim to reengineer some of its components to provide greater scalability by using less number of computational assets.

The amount of land irrigated in the US is approximately the same as their farmers used to irrigate ten years earlier, but the important thing is water they are using nowadays for this purpose is quite less than previously used. They are growing plenty of fruits, vegetables, nuts, and whole grains that fulfill their inhabitant's requirements whole year. Two types of irrigation traditional technologies have been used in the US since 2013 [8]. First one is used in the gravity systems; it makes up 35 to 42% of irrigation systems in the United States. It delivers water from its source to a crop area by flooding through land-forming measures, including canals, waterways, basins, and furrows. Examples are the furrow system controlled flooding systems and uncontrolled flooding systems. The second type of irrigation technology is used by the pressure systems. In pressure systems, tubing or pipes are used to pump water, and irrigation is done through an applicator such as a sprinkler or perforated pipe.

China's development has been affected by three major issues regarding agriculture, landscape, and farmers [9]. The solution to these glitches is agricultural transformation. Though this transformation is not so easy and quick, introducing the cloud computing with Internet of Things to their agriculture is going to help them in solving the issue. However, cloud computing, IoT, and SOA technologies, are helping in the they have built huge data involved agricultural harvesting. Cloud computing is linked to IoT, and both

collectively can enhance the agricultural production to solve the matters regarding agriculture, landscape, and farmers.

In India, different traditional methods are designed and applied regionally in India over the past decades to cope up the necessities of their people in a sustainable way. The three irrigation methods that exist in India are diversion channels, small-scale water bodies such as tanks to store rainwater, and wells to collect groundwater. These methods are for small-scale as well as large-scale applications. As the population of India is increased enough, the needs on the water increase for various drives such as irrigation, domestic, hydroelectricity, industrial, mining, and regeneration. However, India has the largest irrigated area in the whole world, and the irrigated area is only about 40% of the cropped area [10]. One of the main reasons for this low irrigated land is the major use of traditional irrigation methods, which leads to low water use efficiency of about 35–40% [11].

The use of traditional methods without the reach of cloud computing and edge computing causes the unstable watering system for the plants. Consequently, a well-organized and judicious watering system is the major intention. During some last decades, irrigation systems with the use of some sensor networks with different IoT approaches are initiated which basically provides the solution but still they need some improvements. Table 1 shows their water-saving percentages, techniques used by them, and the sensors used by them.

In 2008, Bernard used the rain sensor and estimated the eminence of pasture with and without the sensor. He tried to figure out irrigation water use. He experienced common Bermuda grass to achieve 34% water saving. Xiao et al. [13] self-designed the sensor network for the irrigation system, and they achieved water saving of about 65.22%. Dukes [20] described that water saving of about 40% to 70% can be achieved by using smart controllers but for real-world scenarios of bigger fields; this value can be lessened to 10% [19].

Gutiérrez et al. [14] designed and tried to implement a mechanized irrigation system to use water efficiently. They used a wireless network of some sensors to manage water saving of about 90% as compared to conventional irrigation methods. Similarly, Kumar et al. [5] presented a similar work in the same year and Parameswaran and Sivaprasath [6] and Rawal [16] latterly introduced a few similar sensor-based solutions. Nelson in 2015 used a few sensor data such as temperature and soil moisture and WSAN to automate the irrigation process with decreased water consumption. Saab et al. [17] tried and thrived an on-field survey of a smart phone irrigation setting up. He investigated and tested that application in Mediterranean environments achieved 25% of water saving. Recently, another input to these contributions was made by Saqib (2020), i.e., a network system for the HC12 module is intended to improve the communication range.

3. Architecture of the Proposed System

The anticipated irrigation system is entrenched with the potential smart decisions taking capability to water plants by considering the factors such as crop type, soil type, climate type, temperature, humidity, and soil moisture. Ontology is implanted to query about the decision for a particular plant type, climate type, and soil type, while remaining factors such as temperature, humidity, and soil moisture are sensed by our sensor network. Final decision for watering plants or not relies 50% on the ontology result, and the other 50% is based on our trained machine learning model. The smart architecture of our watering system is given in Figure 1.

Our proposed architecture of IoT has four layers, application layer, processing layer, transport layer, and the perception layer, rather than basic IoT architecture which consists of three layers (application layer, network layer, and perception layer). The perception layer is known as the physical layer, which means it has sensors for assembling data. It senses temperature, soil moisture, and humidity from air. The transport layer is the source of transferring sensed data collected previously to the processing layer through networks such as wireless, 2G, 3G, and LAN. The processing layer stores, scrutinizes, and processes huge amounts of data coming from the transport layer. It utilizes technologies such as databases, cloud computing, and edge computing. The application layer is for providing application-specific services to the end user. Our system deals with the sensors, GSM module, edge server + IoT server, and additionally an android application. These are the perception layer, transport layer, processing layer (cloud computing and intelligent computing), and the application layer, respectively.

3.1. Sensor Data. At first, data are gathered by the sensors as presented in Figure 2. Soil moisture, humidity, and temperature data are collected in this phase. The perception layer has all sensors, actuators, and the microcontroller. Rest is the part of remaining three layers. Transport and processing layers collectively provide schedule for watering crops, their supervision, and other suggestions. After gathering the data, the next stage is to accumulate data at data centers for analyses.

The detailed design inspection of the physical components used is presented in the figure. All the components are with no trouble available in the market and cheap also. So, the device to be deployed in the real environment can be made easily available. This implantable device has the layer of sensors used, i.e., humidity, light, and moisture sensors. The microcontroller fixed in the Arduino board receives the analog signals from these sensors, and after every 30 seconds, these values are transferred to the data center through GSM module SIM808. The final results from our decision-making process can be visualized by the user all the way through an android application, after which the user can direct our system's actuators, and finally, water is released from the valve or closed.

The next section briefs the working of our ML smart decision system deployed at the IoT server which speedily timetables the watering plan for plants. This setting up also evolves the soil type, climate type, and crop type. In our smart system, ontology inhabits in these parameters for better competence and precision. By means of these technologies, we have prepared our system to be fully functionally automatic. The subsequent section describes the semantic knowledge base for our smart irrigation system.

Table 1: Sensor-based solutions.

Work year	Sensor occupied	Technique/methodology	Water saved
Cardenas-Lailhacar et al. (2008) [12]	Rain sensor	Soil moisture sensor system	34%
Xiao et al. (2010) [13]	Self-designed wireless sensor	WSN	65.22%
Gutiérrez et al. (2014) [14]	Soil moisture sensor VH400 Temperature sensor DS1822 Temperature sensor (LM35)	WSN and GPRS	90%
Kumar et al. (2014) [5]	Humidity sensor (CLM53R) Soil pH sensor	WSN and XBee-based communication	No result
Nelson et al. (2015) [15]	Soil moisture sensor	WSAN with cloud platform	72%
Parameswaran and Sivaprasath (2016) [6]	Soil moisture sensor	Drip irrigation with IOT	No result
Rawal (2019) [16]	Soil moisture	Sprinklers with IOT	1000 m^3/ha
Saab et al. (2019) [17]	Blueleaf tool	DSS with IOT	25.7%
Saqib et al. (2020) [18]	Soil moisture sensor	WSNs	No results
Grady et al. (2019) [19]	Prototype/model	Edge computing	No results

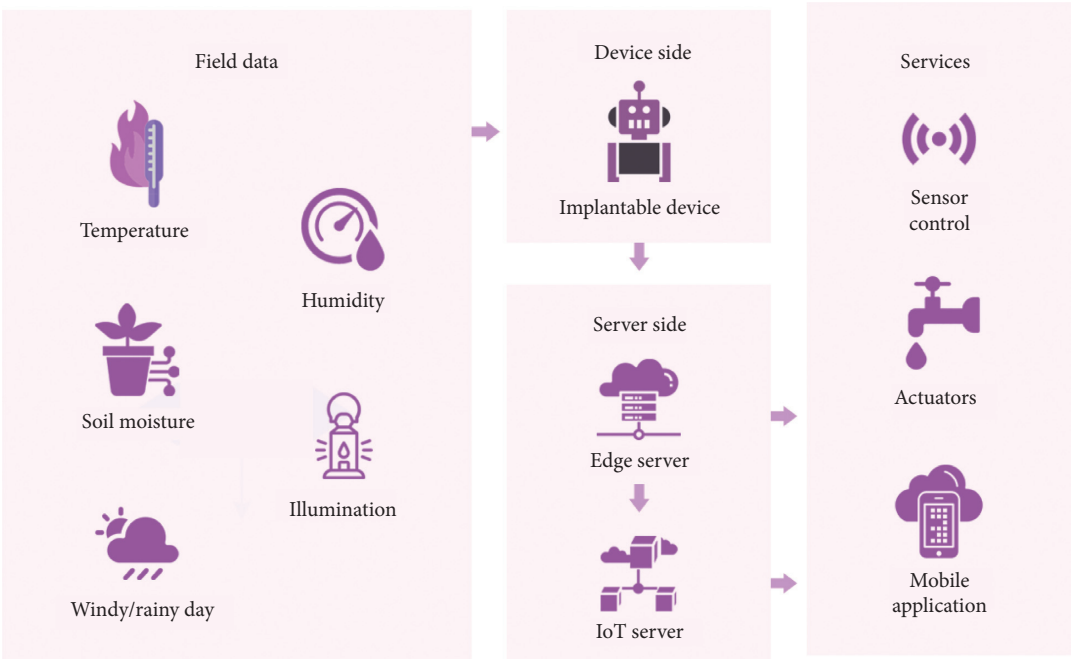


Figure 1: Proposed architecture for smart irrigation.

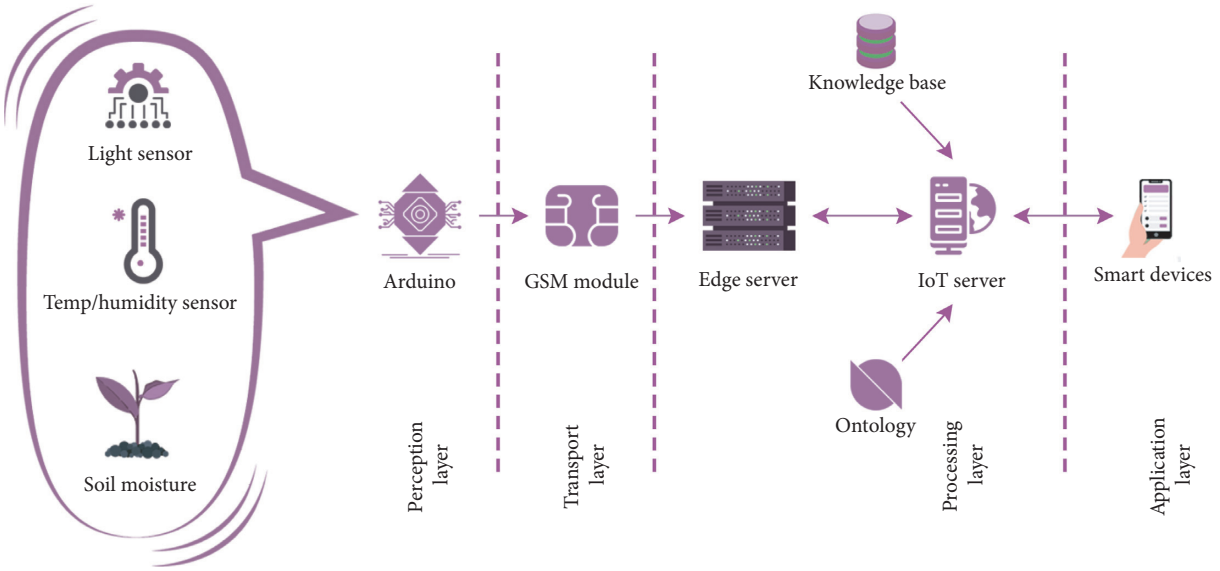


Figure 2: Hardware design for the integrated system.

3.2. Semantic Knowledge Base. The semantic data model (SDM) is designed for incorporating and handling of the real-world data. In the semantic data model, the logical levels are applied for the categorizing of concepts and evaluation of the information. On the basis of the results extracted from propositional logic systems set in the ontology, one can make a smart decision.

There are concepts in our ontology to make prediction of the level of water need on the basis of crop type, climate type, and soil type. These parameters collectively constitute the structured data that why we can query decision on their bases from the given ontology. The sensed data and the decision resulted in SPARQL (RDF query language) together comprise the full ground vital to make a watering system run. For instance, if, due to the climate type and soil strata type, a particular plant requires water, it would be contingent to water the plant. This action of watering crops is the consequence of the actuation that is performed on the valve. Likewise, it could be turned off as directed by the field specification.

Sensor data are pulled together at different levels of a large area. The observed properties or the sensed data such as temperature, humidity, and luminance are measured at the yard level, while soil moisture (superficial and deep) is measured at the quadrant level. These data in the form of RDF and the desired knowledge (crop species, climate types, and soil types) from ontology are sent to the control agent. The control agent also receives data about plant requirement for quantity of water in specific soil texture.

The ontology on which our system depends is vast and complex due to the wide range of factors/features engaged in taking decision for watering plants or not (Appendix A). An abstract view of ontology is shown in Figure 3. There are different climate zones of Pakistan, and they are distinguished into four different types such as highland zone, arid zone, lowland zone, and coastal zone. As the humidity level is diverse in diverse areas, irrigation in these climatic zones has wide-ranging water needs.

In addition to temperature and humidity, another feature, soil type, also influences the level of water need to be given. The clay which is known as well-drained like loamy soils is the excellent soil type for wheat [21]. There are four to five different types of soil considering their structure and texture. In the same way, each crop has its some specific water needs as some require more water such as sugarcane and rice than others such as wheat and cotton.

The architecture of our decision support system is shown in Figure 4. The information about crop types is giving the watering requirements of the crop by utilizing plant ontology. Then, data sensed from pasture/crop land, soil type, and climate type is used for depiction on the actual watering supplies for the field. Water instructions or suggestions will be shown as recommendations on the mobile phone via an android app, and as a resultant of a button click from the farmer's smart phone, actuations will be executed on the valves positioned in the field.

3.3. Used Analysis Technique. Water requirement level can be predicted by any machine learning approach such as random forest, decision trees, KNN, naive Bayes, and support vector machine as all of these are classification dilemma-handling algorithms. The modeling practice we are using lies underneath supervised machine learning, known as KNN (with $k = 5$). It uses the whole dataset to predict an unseen data instance. It searches through the whole dataset to find "k" number of neighbors which are the most close neighbors to that data instance. This is done by actually finding the correspondence between the instance data with the whole dataset, where "k" is the number of neighbors found closer to instant data. If the value of "k" is set to 3, then three most similar neighbors will take part in assigning class label to instant data. It then allocates the most common class label (among those k-training instances) to the test data. Shemim et al. utilized three feature selection algorithms, CBFS, FPRS, and KFRS, for the dataset, and then KNN is used to classify featured classes [21]. Bzdok et al. also discussed about supervised learning algorithms including KNN in 2018 [22].

3.3.1. Algorithm for KNN

Step 1: calculate the Euclidean distance between new data X (4 features involved to predict the resultant class, A, B, C, and D) and each existing point P_n in the input dataset S:

$$\text{Euclidean distance} = \sqrt{(XA - PA)^2 + (XB - PB)^2 + (XC - PC)^2 + (XD - PD)^2}. \quad (1)$$

Step 2: choose the value of "k," i.e., no of nearest neighbors to new data X:

$$k = 5. \quad (2)$$

Step 3: count the ballots of all the "k" neighbors to predict the class of test data X.

Step 4: assign that class to the test data X, which won more votes.

4. Application of the Proposed Architecture

Our system to be implemented uses an Arduino UNO (ATmega328P) controller. The data sensed by the sensors (perception layer) are received by the microcontroller, they are transferred to the edge server (1st processing layer) via GSM SIM808 (transport layer), in which basic scrutiny occurs, and just the immediate data required to predict the resultant water level are transferred to the main IoT server

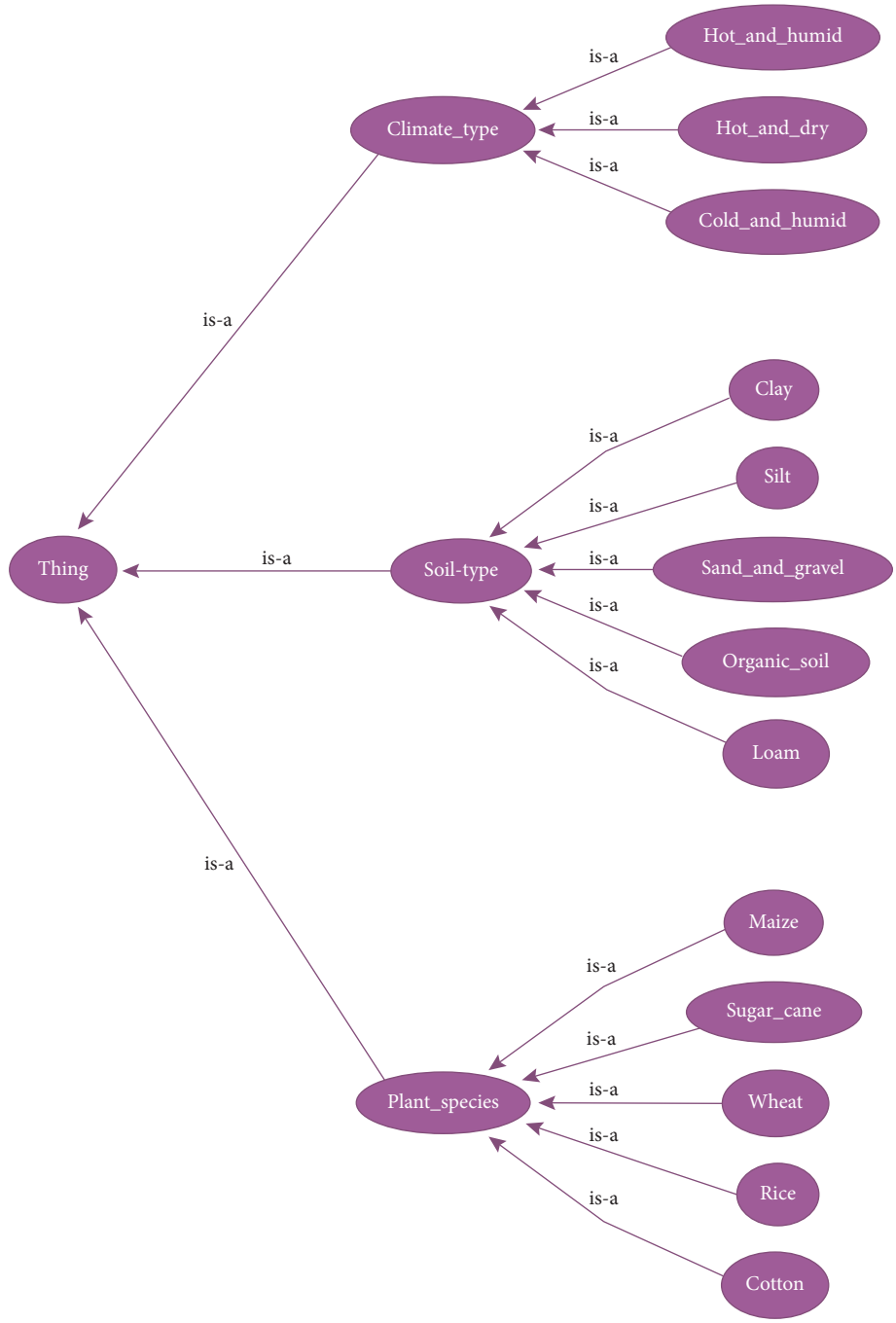


Figure 3: Ontology: an abstract view.

(2nd processing layer) where our trained machine learning model is deployed. This model, after detailed analysis, tells the rank of water need for a field. The following section elaborates the hardware setting.

4.1. Hardware Setting for the User. Embedded sensors used in the IoT-based system are the source of sensing inventively and cost-effectively, and they can record and analyze real-time data (Sarwar, Bajwa, Ramzan et al. 2018 and Munir, Bajwa and Cheema 2019) [23]. The proposed smart IoT system as shown in Figure 5 employs some sensors to gather data from the environment, and a GSM module SIM808 is used to transfer the values to the edge server. A data SIM card is inserted in it to get facilitated by the real-time data transportation. As we can see in Figure 6, a hygrometer sensor is used for soil moisture, while for the moisture from air, AM2302 DHT22 (temperature/humidity) sensor is used. Their details are described in the following.

4.1.1. HL-69 Soil Hygrometer Sensor. For the detection of the humidness of the soil, we used HL-69 soil hygrometer moisture sensor. The basic purpose for using the HL-69 soil

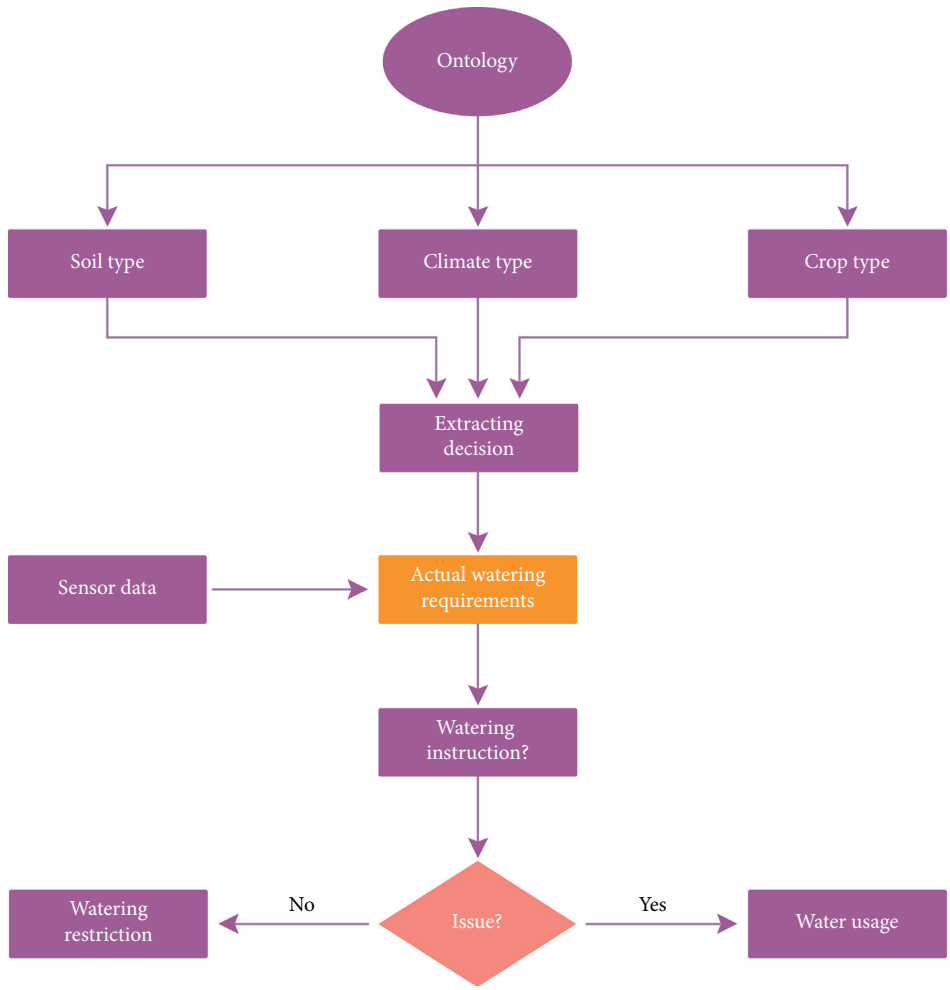


Figure 4: Inference rule schema.

hygrometer moisture sensor is to provide better reading than other soil moisture sensors. This sensor is used for real-time monitoring soil moisture of plants in a tunnel farm. The voltages of the sensor output change accordingly to the water content in the soil. There are some key factors of HL-69 soil hygrometer sensor. If soil moisture is greater, then the output voltage decreases, but if the soil is dry, then the output voltage increases. The hygrometer soil moisture sensor provides an analog signal as an output which has to be converted to digital by Arduino. This sensor includes two pieces: one is an electronic board and another one is two pads that detect the water content. LM393 comparator chip is located on the electronic board. The electronic board of the HL-69 soil hygrometer sensor has a fixed bolt hole used for easy installation. It contains two lights: red and green; red light shows the power indicator, and the green light shows the digital switching output indicator.

4.1.2. AM2302 DHT11 Sensor. The DHT22 sensor is a common temperature-humidity sensor that is used to determine temperature and humidness in air. The DHT22 sensors are made up of two parts: a humidity sensor and a thermistor. There are some key factors of the DHT22 sensor which are as follows: the cost of the DHT22 sensor is low. DHT22 sensor is good for 0–50% temperature readings with 2–5% accuracy and a humidity range from 20 to 80% with 5% accuracy. The I/O voltage for the DHT22 sensor is between 3 V and 5 V. While requesting data, the maximum current use during conversion is 2.5 mA. DHT22 sensor contains 4 pins with 0.1 spacing between them. The body size of the DHT22 sensor is approximately 15.1 mm *25 mm *7.7 mm.

4.1.3. BH1750 FVI Light Sensor. BH1750 is a common digital light sensor that can determine the light intensity. BH1750 is a calibrated digital light sensor, and it can measure even small traces of light and can convert it into a 16-digit numeric value. It is commonly used in mobile phones to exploit the screen brightness based on the environmental lighting. BH1750 measures the light intensity in the range of 0 to 65,535 lux (L). In the smart irrigation system for tunnel farming, we used the H-resolution mode of the BH1750 sensor. There are some key factors of BH1750 sensor which are as follows: the chip of the BH1750 sensor is BH1750FVI. The power supply of the BH1750 sensor is 3.3 V to 5 V. The BH1750 sensor is a built-in 16 bit AD converter that converts detection of light into a 16-digit numeric value.

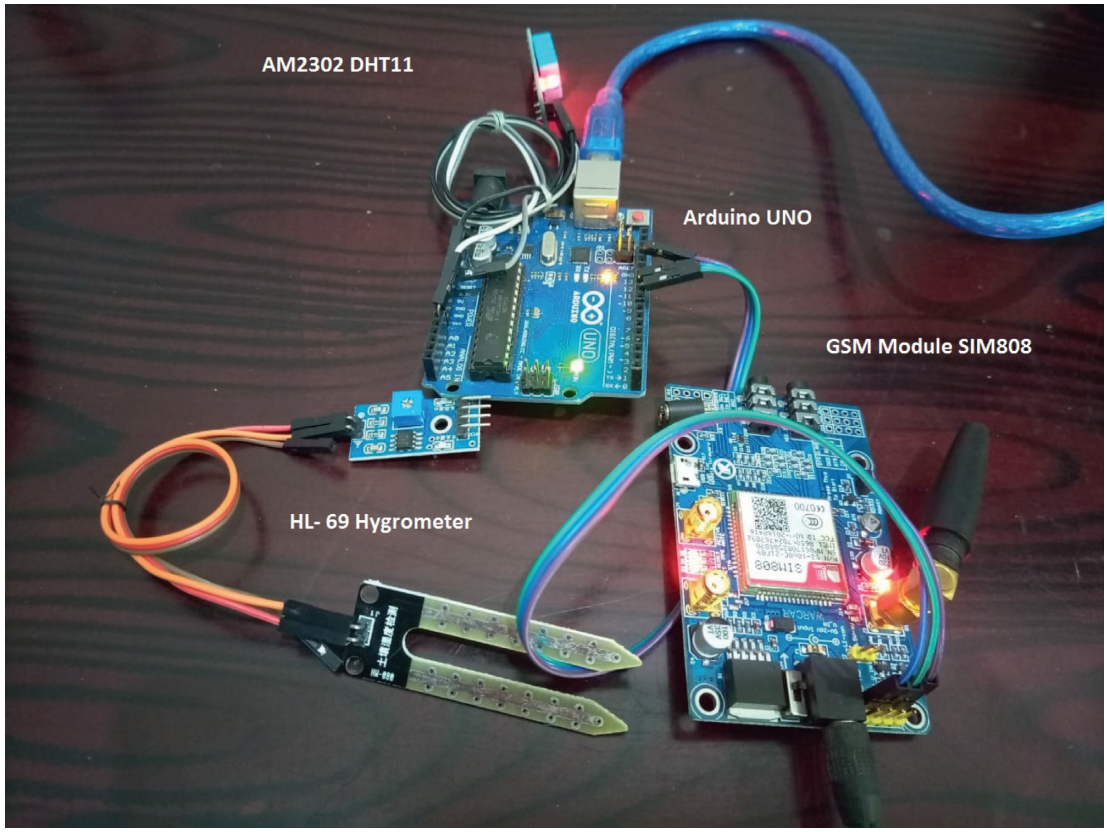
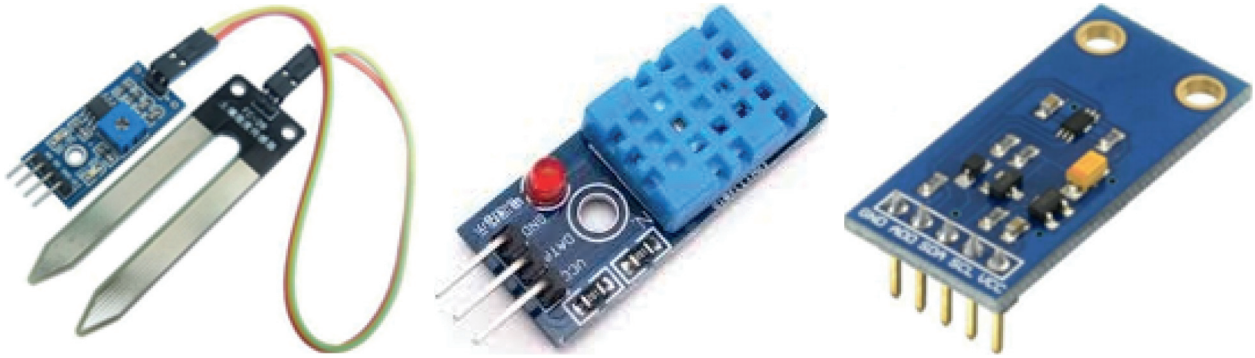


Figure 5: Hardware prototype.



HL-69 hygrometer soil moisture sensor

AM2302 DHT22 temperature-humidity sensor

BH1750 light sensor

Figure 6: Sensors used.

The range of light intensity in the BH1750 sensor is 0 to 65,535 lux. The size (L*W) of the BH1750 sensor is approximately 3.2 cm*1.5 cm.

5. Results and Discussion

The proposed watering system for tunnel farming is so smart that develops and employs the assistance of true decision-making capability of machine learning. The architecture and the hardware details of the system are given in the preceding section. All the sensors (temperature and humidity, light sensor, and the soil moisture sensors) were deployed to the actual field to analyze the reaction of the proposed system. The data transferred to the edge server through the GSM module and through an Android application whereas the results can also be seen by a farmer. A user can then perform some actuation to open or close the valve.

5.1. Preparing the Training Dataset. The system is completely automated as the sensor data receiving from the field are processed according to our trained model of machine learning, i.e., trained by the characteristic sensor values shown in the following. Table 2 shows five classes: highly

needed, needed, average, not needed, and highly not needed for various levels of soil moisture sensed by the HL-69 hygrometer sensor and temperature and humidity sensed by AM2302 DHT22. The output of a HL-69 hygrometer varies from 0 to 870, while the humidity level of the AM2302 DHT11 sensor varies in the air from 20 to 80%, and its temperature value ranges from 0 to 50.

Here is our sampled training dataset shown in Figure 7 based on our rules set in Table 2, which we have provided to our machine learning algorithm to predict water needs for the given crop types.

5.2. Training of the KNN Model. We have implemented the code in Anaconda, created for python programs, and have trained our model on the fact that, on a particular temperature, water requirements of different crops, which we are taking into consideration, can be given.

Rice > sugarcane > maize > cotton > wheat.

This general rule can be elaborated more specifically by identifying ranges for temperature, humidity, and soil moisture for all the given types of crops to recognize its class. Here is the rule summary in Table 3.

In Table 3, labels "HN," "N," "A," "NN," and "HNN" are second hand for classes highly needed, needed, average, not needed, and highly not needed correspondingly. Likewise, working rules for watering considering climate and soil are given in the following.

Sand and gravel > clay > silt > loam > organic soil.

Hot and dry > hot and humid > cold and humid.

5.3. Deploying the Trained Model. Our trained model has been developed using Scikit-learn. To make it available to production and to make it useful for end users so that they could extract real values from it, we have deployed it. In this regard, we need to have three components shown in Figure 8.

The developed and trained model for predicting the water level as a "model.pkl" file is ready, and model evaluation is provided in the results section. The web service we have used for this purpose is the Flask API. Lastly, we need cloud as a service provider, and Heroku server fulfilled our requirement in this proposed system.

5.4. Implementation Using Edge Computing. In the process of deploying the trained model through Flask API, we actually define routes to where an HTTP request handles. One route is for handling one HTTP request. The data are travelled in the system from one side (perception layer) to other (edge server).

The piece of code in Figure 9 is responsible for sending the sensor data from the sensor-Arduino side (i.e., perception layer) to edge server Firebase (i.e., processing layer). Second last line of the code is establishing a link to which data (temperature, humidity, soil moisture, and phone no.) are transferred. These data are received by our edge server by a route "/submit" defined in the application of Flask API as shown in Figure 10. Whenever data from sensors send to the established link of HTTP request, it triggers the following piece of code to run. In this piece of code, the sensor values and the SIM card no. (phone no.) are inserted in our database server (Firebase).

Table 2: Classes of sensor data.

Soil moisture (%)	Temperature (°C)	Humidity (%)	Class
<30	>45	<30	Highly needed
30–45	35–45	30–45	Needed
46–60	25–34	46–60	Average
61–80	20–24	61–80	Not needed
80–100	<20	>80	Highly not needed

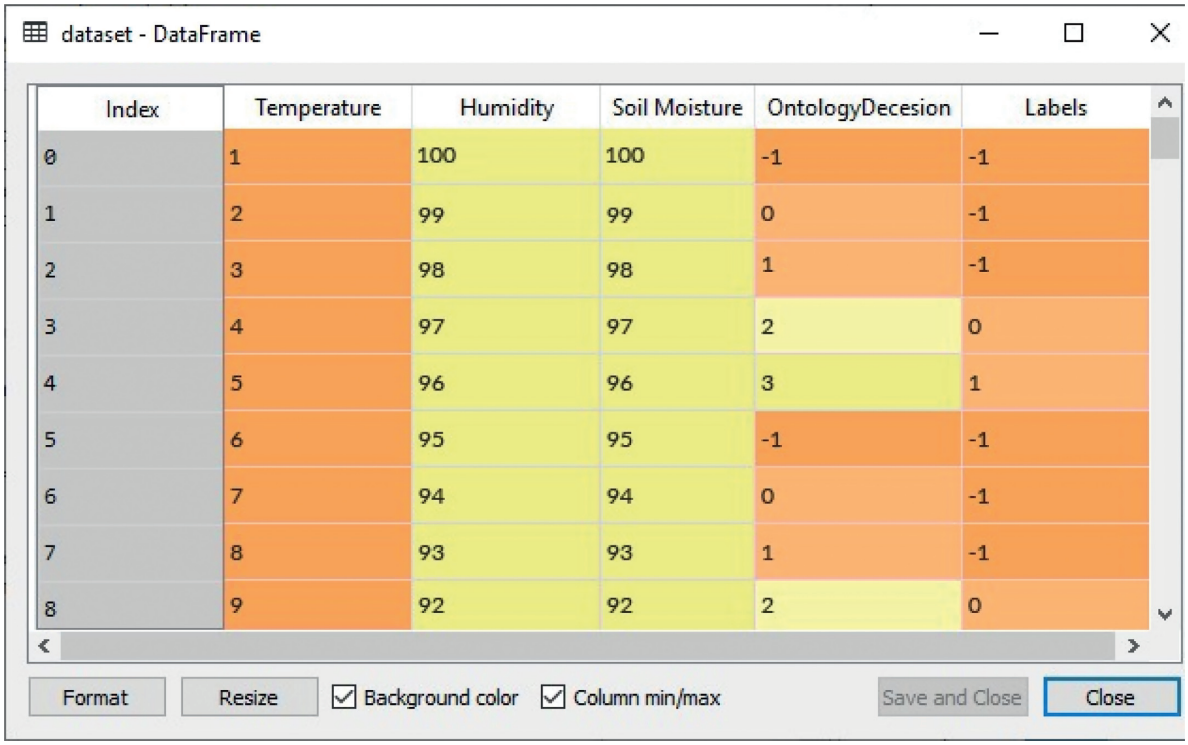
As we can see in Figure 11, each phone no. is representing a different device. Any data coming from a particular device are stored under the hierarchy of its phone no. Each record under a device has a key value associated with it, which actually contains the sensor data values. Whenever a particular record arrives at the edge server, its key value stores in the parameter "latestKey" under its phone no. When another entry of record happens, its key value replaces the previous one. In this way, our database is designed to have the record of most recent data entered in it.

Firebase can be omitted from the system, and data can directly be sent to the Heroku server (IoT server), i.e., cloud computing. That is really a bad practice due to overburden of the IoT server with useless data. Since sensors are sending each and every instant value to the IOT server and IOT server is responsible for scrutiny of data coming from a device (which means three sensors values every 30 seconds), and then applying model to predict value for water requirement. It will definitely affect the speed and efficiency of the system. This is the main concept of introducing edge computing.

The piece of code in Figure 12 is triggered by the smart mobile/tab when the user clicks to predict results for water requirements. It utilizes the trained model to predict the water requirements and return the value to the server. This value is sent to the user, and he/she can see the result on his/her phone via the app.

5.5. Android Application. An android platform is provided to the farmers. The input parameters (crop type, climate type, and soil type) are put on view in a dropdown, and users can select from these and can send the command to the device implanted to the field.

Codes for crop types, soil types, and climate types are transferred from the mobile app interface in Figure 13(a) to the server to which ontology is attached. Decision extracted from the ontology section along with the sensor values then reaches the main IoT server where our machine learning algorithm is installed. Our training dataset also contains the encoded values for labels for different classes which are converted to the text (class label) at the front end in the android app as in Figure 13(c). These codes are shown in Table 4.



Index	Temperature	Humidity	Soil Moisture	OntologyDecesion	Labels
0	1	100	100	-1	-1
1	2	99	99	0	-1
2	3	98	98	1	-1
3	4	97	97	2	0
4	5	96	96	3	1
5	6	95	95	-1	-1
6	7	94	94	0	-1
7	8	93	93	1	-1
8	9	92	92	2	0

Figure 7: Sampled training dataset.

Table 3: Working rules for different crops.

Rule no.	Temperature	Humidity	Soil moisture	Ontology decision	Class
1	T >50	H <20	SM <20	HN	HN
2	T >40	H <40	SM <40	HN	N
3	T >30	H <60	SM <60	HN	A
4	T >20	H <80	SM <80	HN	NN
5	T <20	H >80	SM >80	HN	HNN
6	T >57	H <20	SM <10	N	HN
7	T >40	H <30	SM <30	N	N
8	T >35	H <40	SM <40	N	A
9	T >30	H <60	SM <60	N	NN
10	T <30	H >60	SM >60	N	HNN
11	T >57	H <20	SM <10	A	HN
12	T >40	H <30	SM <30	A	N
13	T >35	H <40	SM <40	A	A
14	T >30	H <60	SM <60	A	NN
15	T <30	H >60	SM >60	A	HNN
16	T >50	H <30	SM <40	NN	HN
17	T >40	H <40	SM <60	NN	N
18	T >30	H <60	SM <80	NN	A
19	T >20	H <90	SM <100	NN	NN
20	T <20	H >90	SM >100	NN	HNN
21	T >50	H <30	SM <30	HNN	HN
22	T >40	H <50	SM <60	HNN	N
23	T >30	H <60	SM <80	HNN	A
24	T >20	H <80	SM <90	HNN	NN
25	T <20	H >80	SM >90	HNN	HNN

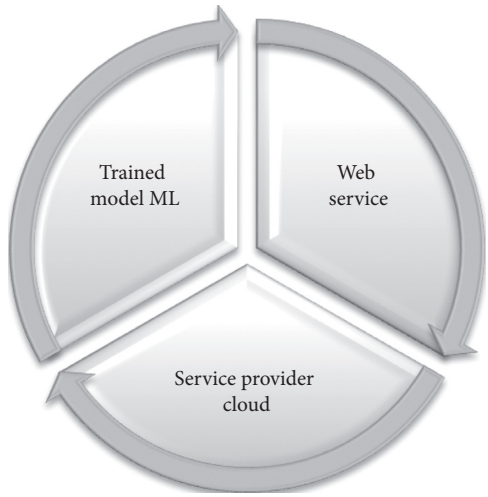


Figure 8: Components required in deployment.

SIM808_HTTP-POST | Arduino 1.8.9

File Edit Sketch Tools Help

Verify

SIM808_HTTP-POST §

```
String url = "https://irrigate-api.herokuapp.com/submit?";  //URL for HTTP-POST-REQUEST
String temp;
String humi;
String soilmoisture;
String phone;
 void gsm_sendhttp() {
  mySerial.println("AT+HTTPINIT");
  runsl();
  delay(100);
  mySerial.println("AT+HTTPPARA=CID,1");
  runsl();
  delay(100);
  mySerial.println("AT+HTTPPARA=URL," + url+"temp="+temp +"&humi="+ humi + "&soilmoisture=" + soilmoisture + "&phone=" + phone);
  runsl();
```

Figure 9: HTTP request sending to store sensor data.

```
@app.route('/submit', methods=['POST'])
def submit():
    dump(request)
    if request.values and len(request.values) == 4:
        data = {
            'temperature': request.values['temp'],
            'humidity': request.values['humi'],
            'soil_moisture': request.values['soilmoisture']
        }
        print('data ===> ', data)
        result = firebase.post('/sensor-data-collection-7d34e/data/' + request.values['phone'], data)
        firebase.put('/sensor-data-collection-7d34e/data/' + request.values['phone'], 'latestkey', result)
        return json.dumps({'status': 'true', 'message': 'Data Saved!'})
    else:
        return json.dumps({'status': 'false', 'message': 'Invalid request!'})
```

Figure 10: Route defined for inserting values to the database.

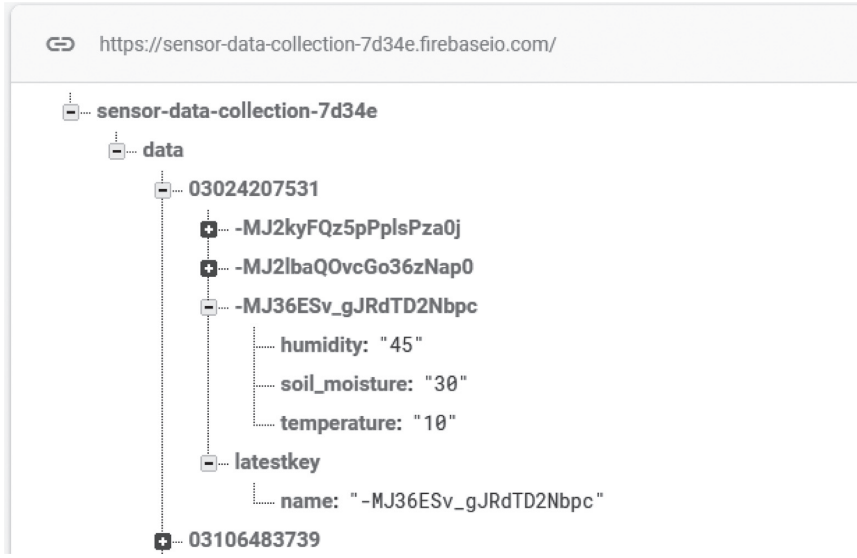


Figure 11: Edge server handling data from sensors via the HTTP request.

```
@app.route('/predict', methods=['POST'])
def predict():
    dump(request)
    if request.values and len(request.values) == 4:
        crop = request.values['crop']
        climate = request.values['climate']
        soil = request.values['soil']
        phone = request.values['phone']
        latestkey = firebase.get('/sensor-data-collection-7d34e/data/' + phone + '/latestkey', 'name')
        humidity_result = firebase.get(
            '/sensor-data-collection-7d34e/data/' + phone + '/' + latestkey,
            'humidity')
        soilmoisture_result = firebase.get(
            '/sensor-data-collection-7d34e/data/' + phone + '/' + latestkey,
            'soil_moisture')
        temperature_result = firebase.get(
            '/sensor-data-collection-7d34e/data/' + phone + '/' + latestkey,
            'temperature')
        model = pickle.load(open("./model/model.pkl", "rb"))
        prediction = model.predict(
            [[humidity_result, soilmoisture_result, temperature_result, crop, climate, soil]])
        print(prediction)
        return json.dumps({'status': 'true', 'message': str(prediction[0])})
    else:
        return json.dumps({'status': 'false', 'message': 'Invalid request.'})
```

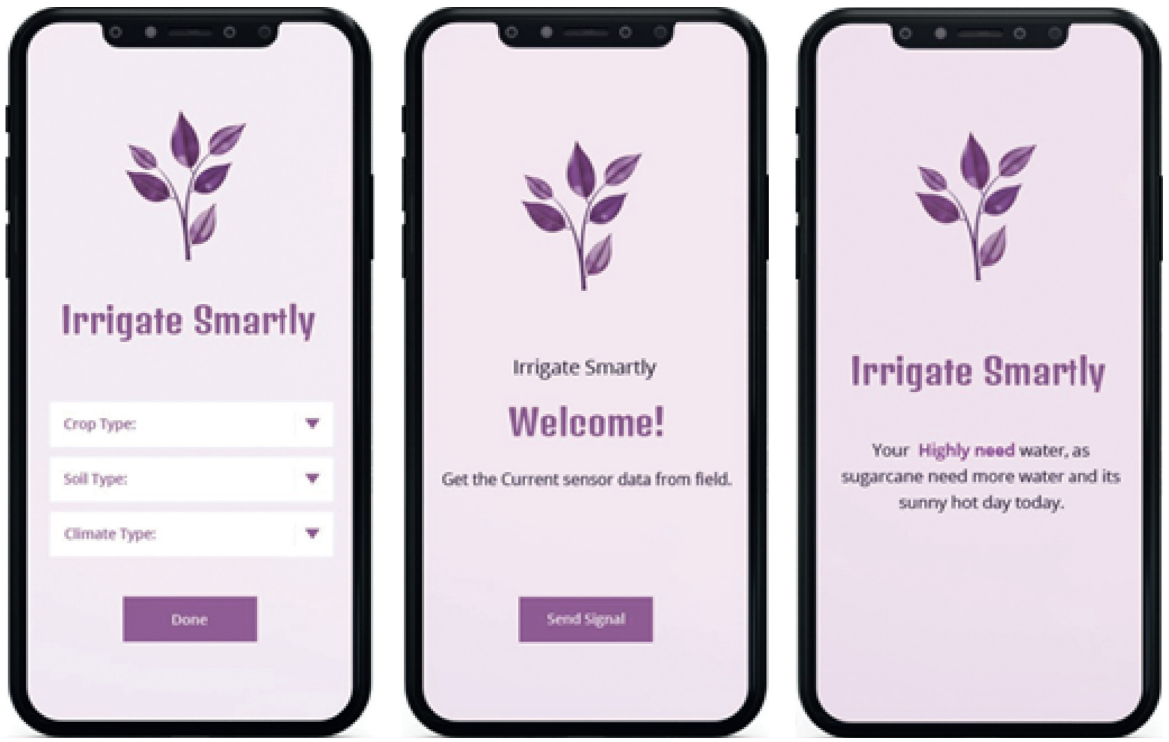
Figure 12: Data transfer from edge server to IoT server with prediction results.

For example, we choose, from the dropdowns in the user input screen shown in Figure 13(a), sugarcane as a crop type, hot and dry as a climate type, and loam as a soil type. After clicking the button "Send" from the Figure 13(b) interface, the sensor readings come across to the server.

The values for humidity, temperature, and soil moisture and the encoded result for soil type, climate type, and crop type values are considered by our trained ML model to recognize the watering need for the specific crop. So, with the 50% result coming from ontology and the sensor values, our system foretells to water the crops and displays a note on the farmer's mobile screen as shown in Figure 13(c).

The highlighted text "Highly need water" is mainly the output of our machine learning algorithm already discussed in the previous section. As shown in Table 4, our training dataset holding the labels ranges from −1 to 3. These are effectively the degree of water necessity to a specific plant at specific time.

5.6. Performance Evaluation. We have performed tests on our sample data, which we have obtained randomly from about 500 instances. We provided these instances to train our KNN model for the proposed system to forecast class



(a) (b) (c)

Figure 13: Android app different interfaces (a, b, and c).

Table 4: Codes for class labels.

Class	Code
Highly needed	3
Needed	2
Average	1
Not needed	0
Highly not needed	−1

labels. We used two statistical measures to estimate the performance of our KNN model, i.e., precision and recall. Figure 14 shows the accuracy report of the results of the KNN model providing $k = 5$.

The accuracy report of the trained model is given in Figure 1 and also presented graphically. Predicted results for class label "Needed" are lacking in precision. This performance is dependent on the "k" value that is the no. of nearest neighbors involved in the predicting class. Figure 15 shows the confusion matrices without normalization and with normalization.

To increase accuracy, we should choose the "k" value precisely. As per general rules for the KNN algorithm, the value of "k" for the problem of two classes should be an odd value, and for more than two classes, the "k" value should not be the multiple of the number of the resultant classes. As in our case, we have five labeled classes to be predicted, so we will choose "k" accordingly. In order to choose a suitable value for "k", we have to plot "k value versus mean error" graph to identify the error trend. So, we plot it by using "matplotlib.pyplot" in Figure 16.

As we can see, the mean error initially increases up to 0.5 as the "k" value increases, but there is a sudden fall which occurs after that to the value of 0.3 when the "k" value reaches 10 to 11. After that, rise in error rate starts, and it continues to increase with the increase in the "k" value. It means that our "k" value should be "11" that is the maximum value of 'k' for which mean error remains lowest. So, for the value of $k = 11$, we again find the accuracy report to check if the performance of our model is getting better or not.

Figure 17 shows the significant improvement in the performance of the model as accuracy rate increases when we set the "k" value to 11. The precision value for class "Not Needed" is increased from 0.33 to 1.0 which means accuracy rate increases from 33% to 100%. Similarly, for class "Average," the precision value increases from 0.25 to 0.33 which means that accuracy rate increases from 25% to 33%. This is how tuning of the model can be possible. By tuning training data values and by adjusting the "k" value, we can have a better model for our system. This is the main reason why we choose KNN algorithm for our proposed system.

	Precision	Recall	F1-score
−1	0.8	1	0.89
0	0.33	0.2	0.25
1	0.25	0.33	0.29
2	0	0	0
3	1	1	1
Accuracy			0.45
Macro avg	0.48	0.51	0.48
Weighted avg	0..42	0.45	0.43

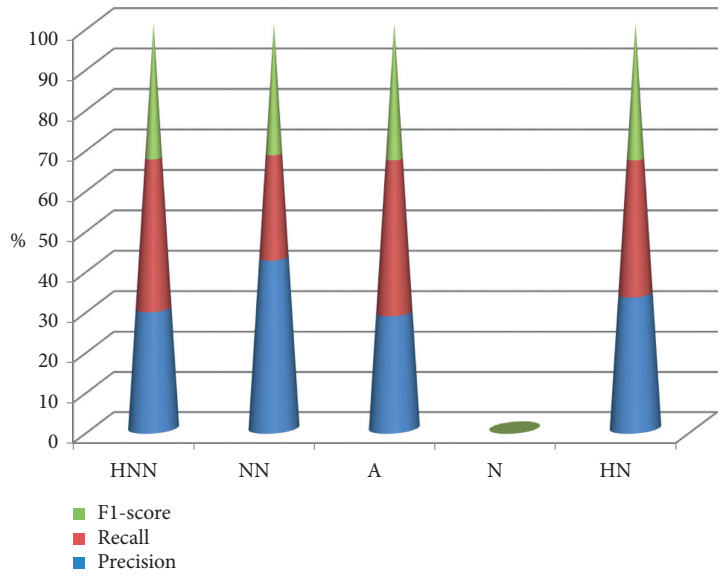


Figure 14: Accuracy report with "$k = 5$."

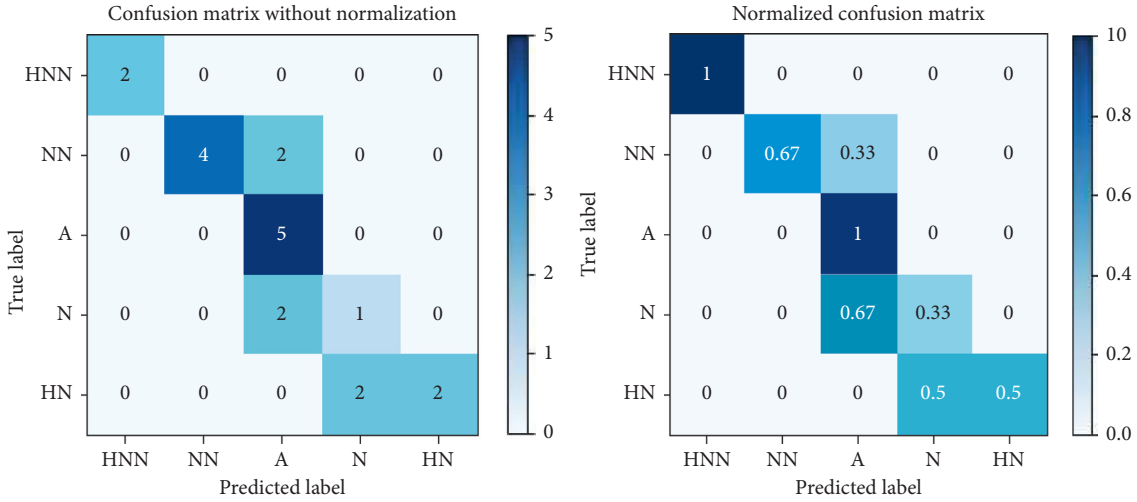


Figure 15: Confusion matrices without normalization for "$k = 5$."

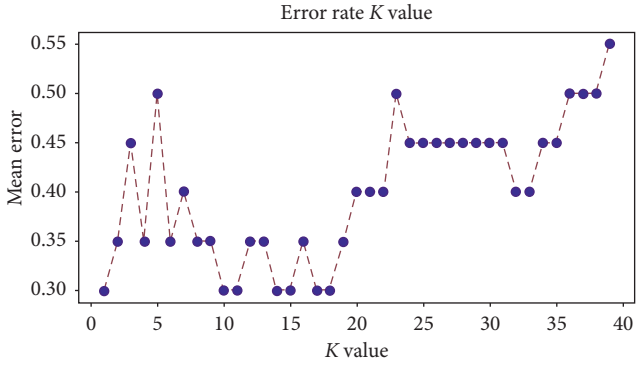


Figure 16: Error rate with respect to the k value.

	Precision	Recall	F1-score
−1	0.8	1	0.89
0	1	0.5	0.67
1	0.33	1.00	0.50
2	0	0	0
3	0.83	0.83	0.83
Accuracy			0.65
Macro avg	0.59	0.67	0.58
Weighted avg	0.64	0.65	0.61

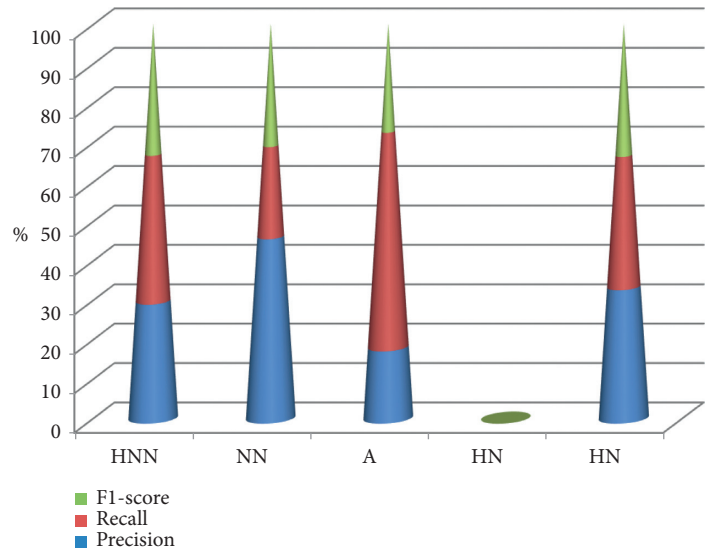


Figure 17: Accuracy report with "$k = 11$."

6. Conclusion and Future Work

Our proposed solution for smart irrigation constitutes three modules: first module is the sensor network, which is required to sense parameters influencing the water need. We have used sensors DHT22, light sensor BH1750, and HL-69 hygrometer to sense temperature, soil moisture, light, and humidity in air. In the third module, we use edge and main IoT servers to transfer and receive data via HTTP requests. In the second module, we applied KNN on the sample dataset to train the model and used it for efficient decision-making of water requirements. Our trained model classifies the input into five possible classes based on input values such as highly not required, not required, average, required, and highly required. We have fully implemented the proposed system in Anaconda.

Currently, our system employs KNN for decision-making, but other intelligent data-extracting techniques can also be used for decision-making. So, the presented irrigation system can reproduce in future by using other decision-making techniques such as random forest. Moreover, the edge computing architecture can be further improved by making the edge server responsible for processing data and depicting the result from the machine learning algorithm. In other words, the trained model for KNN can be deployed at the edge server so that nearby devices to a particular edge can get facilitated by that edge server. It will improve latency rate remarkably.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] M. S. Munir, I. S. Bajwa, M. A. Naeem, and B. Ramzan, "Design and implementation of an IoT system for smart energy consumption and smart irrigation in tunnel farming," *Energies*, vol. 11, no. 12, p. 3427, 2018.

[2] H. Sattar, I. S. Bajwa, R. U. Amin et al., "An IoT-based intelligent wound monitoring system," *IEEE Access*, vol. 7, pp. 144500–144515, 2019.

[3] B. Sarwar, I. Bajwa, S. Ramzan, B. Ramzan, and M. Kausar, "Design and application of fuzzy logic based fire monitoring and warning systems for smart buildings," *Symmetry*, vol. 10, no. 11, p. 615, 2018.

[4] B. Sarwar, I. S. Bajwa, N. Jamil, S. Ramzan, and N. Sarwar, "An intelligent fire warning application using IoT and an adaptive neuro-fuzzy inference system," *Sensors*, vol. 19, no. 14, p. 3150, 2019.

[5] A. Kumar, K. Kamal, M. O. Arshad, S. Mathavan, and T. Vadamala, "Smart irrigation using low-cost moisture sensors and XBee-based communication," in *Proceedings of the Global Humanitarian Technology Conference (GHTC)*, San Jose, CA, USA, October 2014.

[6] G. Parameswaran and K. Sivaprasath, "Arduino based smart drip irrigation system using internet of things," *International Journal of Engineering Science*, vol. 6, p. 5518, 2016.

[7] C. Kamienski, J.-P. Soininen, M. Taumberger et al., "Smart water management platform: iot-based precision irrigation for agriculture," *Sensors*, vol. 19, no. 2, p. 276, 2019.

[8] M. Stubbs, *Irrigation in US Agriculture: On-Farm Technologies and Best Management Practices*, Congressional Research Service, Washington, DC, USA, 2016.

[9] F. TongKe, "Smart agriculture based on cloud computing and IOT," *Journal of Convergence Information Technology*, vol. 8, no. 2, 2013.

[10] A. Narayanamoorthy, "Economics of drip irrigation in sugarcane cultivation: case study of a farmer from Tamil Nadu," *Indian Journal of Agricultural Economics*, vol. 60, pp. 235–248, 2005.

[11] M. W. Rosegrant, X. Cai, and S. A. Cline, "Global water outlook to 2025: averting an impending crisis," International

Food Policy Research Institute, Washington, DC, USA, 572-2016-39087, 2002.

[12] B. Cardenas-Lailhacar, M. D. Dukes, and G. L. Miller, "Sensor-based automation of irrigation on Bermuda grass, during dry weather conditions," *Journal of Irrigation and Drainage Engineering*, vol. 134, pp. 184–193, 2008.

[13] K. Xiao, D. Xiao, and X. Luo, "Smart water-saving irrigation system in precision agriculture based on wireless sensor network," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 26, pp. 170–175, 2010.

[14] J. Gutiérrez, J. F. Villa-Medina, A. Nieto-Garibay, and M. A. Porta-Gandara, "Automated irrigation system using a wireless sensor network and GPRS module," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 1, pp. 166–176, 2014.

[15] N. Sales, O. Remédios, and A. Arsenio, "Wireless sensor and actuator system for smart irrigation on the cloud," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pp. 693–698, IEEE, Milan, Italy, December 2015.

[16] S. Rawal, "IOT based smart irrigation system," *International Journal of Computer Applications*, vol. 159, no. 8, pp. 7–11, 2017.

[17] A. Saab, M. Therese, I. Jomaa, S. Skaf, S. Fahed, and M. Todorovic, "Assessment of a smartphone application for real-time irrigation scheduling in Mediterranean environments," *Water*, vol. 11, p. 252, 2019.

[18] M. Saqib, T. A. Almohamad, and R. M. Mehmood, "A low-cost information monitoring system for smart farming applications," *Sensors*, vol. 20, no. 8, p. 2367, 2020.

[19] M. J. OGrady, D. Langton, and G. M. P. O'Hare, "Edge computing: a tractable model for smart agriculture?," *Artificial Intelligence in Agriculture*, vol. 3, pp. 42–51, 2019.

[20] M. D. Dukes, "Water conservation potential of landscape irrigation smart controllers," *Transaction ASABE*, vol. 55, pp. 563–569, 2012.

[21] M. S. Munir, I. S. Bajwa, and S. M. Cheema, "An intelligent and secure smart watering system using fuzzy logic and blockchain," *Computers & Electrical Engineering*, vol. 77, pp. 109–119, 2019.

[22] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: supervised methods," *Nature Methods*, vol. 15, pp. 5-6, 2018.

[23] M. Safdar Malik, I. Sarwar Bajwa, and S. Munawar, "An intelligent and secure IoT based smart watering system using fuzzy logic and blockchain," *Computers and Electrical Engineering*, vol. 77, no. 1, pp. 109–119, 2018.

Hindawi
Complexity
Volume 2021, Article ID 5542889, 10 pages
https://doi.org/10.1155/2021/5542889

Research Article

An Improved Prediction Model of IGBT Junction Temperature Based on Backpropagation Neural Network and Kalman Filter

Yu Dou

School of Engineering, University of Leicester, Leicester LE1 7RH, UK

Correspondence should be addressed to Yu Dou; yd116@leicester.ac.uk

Received 19 January 2021; Revised 4 February 2021; Accepted 19 February 2021; Published 26 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

With the rapid development of emerging technologies such as electric vehicles and high-speed railways, the insulated gate bipolar transistor (IGBT) is becoming increasingly important as the core of the power electronic devices. Therefore, it is imperative to maintain the stability and reliability of IGBT under different circumstances. By predicting the junction temperature of IGBT, the operating condition and aging degree can be roughly evaluated. However, the current predicting approaches such as optical, physical, and electrical methods have various shortcomings. Hence, the backpropagation (BP) neural network can be applied to avoid the difficulties encountered by conventional approaches. In this article, an advanced prediction model is proposed to obtain accurate IGBT junction temperature. This method can be divided into three phases, BP neural network estimation, interpolation, and Kalman filter prediction. First, the validities of the BP neural network and Kalman filter are verified, respectively. Then, the performances of them are compared, and the superiority of the Kalman filter is proved. In the future, the application of neural networks or deep learning in power electronics will create more possibilities.

1. Introduction

The insulated gate bipolar transistor (IGBT) is the main power electronic energy conversion device and transmission device. It combines the merits of MOSFET and BJT with low drive power and low saturation voltage. The IGBT has become the pivotal supporting technology to alleviate energy shortages and reduce carbon emissions since it is highly efficient, energy-saving, and environmentally friendly. Nowadays, it is widely utilized in communication, rail transit, smart grid, aerospace, electric vehicles, and new energy power generation [1]. Given the significance of the IGBT, it is essential to maintain safety and reliability during the IGBT device's operation.

One of the most important factors affecting the technical progress and development of the IGBT is operating temperature. The high junction temperature resulted from large heat fluxes will significantly deteriorate the performance and reliability of the IGBT device [2]. The electronics prognostics supplied by the NASA AMES Laboratory also pointed out that overheating of the IGBT die is one of the main causes of the failure [3]. In a related survey on the power device's reliability, the failure rate due to junction temperature is as high as 55% and doubles for every 10°C increase [4]. Thus, ensuring the IGBT junction is maintained at a controllable temperature is the cornerstone of keeping the IGBT device's stability. Monitoring IGBT junction temperature during the operation has become the major challenge and top priority at present.

There have already been many approaches to monitor junction temperature, including optical, physical, and electrical methods [5, 6]. The optical method measures the energy change of lattice photons to infer the junction temperature using infrared (IR) cameras [7], IR sensors [8], IR microscope [9], and optical fiber [10]. This method can directly and accurately obtain the junction surface temperature map, but the implementation usually requires expensive instruments and extra modification on the standard module package. The typical physical method is the electrothermal model, which uses thermistors or thermocouples to physically contact the IGBT chip and infer the junction temperature [11, 12]. However, this method's

response time is usually long because of the thermal capacitance of thermistors and thermocouples [5]. Besides, the physical method relies too much on complex physical models. The accuracy of prediction will be greatly reduced once the physical models change for some reason. The electrical method uses temperature-sensitive electrical parameters (TSEPs) such as gate threshold voltage (V_{th}) [13], on-state voltage (V_{CEon}) [14], short circuit current (I_{sc}) [15], and peak gate current (I_{Gpeak}) [16] to calculate or infer the junction temperature. It can be found either in the scientific literature or the device datasheet published by the manufacturer that there is good linearity between TSEPs and junction temperature. A major benefit of using the electrical method is that the junction temperature can be obtained without modifying the standard module package, which is also the main reason it is used more now. This article will focus on the approach of using on-state voltage (V_{CEon}) at a high current.

The conventional V_{CEon} (high-current) TSEP method attempts to build a temperature model using linear equations. However, the relationship between on-state voltage and junction temperature does not perform absolute linearity. There is always an error between the real value and the calculated value [17]. Dong and coworkers recently proposed a new junction temperature prediction model using an artificial neural network (ANN) [18]. They applied the backpropagation (BP) neural network to predict the junction temperature using on-state voltage and collector current. The results are compared with the conventional TSEP method, and the feasibility of the BP neural network is proved. However, because of the inherent data, randomness during the training process, and intrinsic nonidentifiability of the model, the prediction results are prone to strong instability [19]. This article proposes a new method using the Kalman filter based on BP neural network and interpolation to stabilize the prediction. The flowchart of this approach is shown in Figure 1.

This article is organized as follows: Section 2 describes the object of study and simulation environment. Section 3 introduces three methods used to predict junction temperature and their parameter settings. The results are given in Section 4. The validities of the BP neural network and Kalman filter are verified. There is also a comparison between the errors of the three predicting methods in this section. Section 5 discusses the superiority of the Kalman filter prediction and the possibility of improvement. Finally, the conclusion is drawn.

2. Simulation Settings

This article's research object is Infineon IKW75N65ET7 IGBT discrete (650 V, 75 A) and the simulation is based on LTspice. The SPICE model downloaded from Infineon has already included the temperature module. The on-state voltage test circuit is built to measure the on-state voltage V_{CEon} of the IGBT discrete. The test circuit includes current source, power load, IGBT discrete, control system, and voltmeter, as shown in Figure 2. The junction temperature extraction can be carried out without modifying the standard package even under the real experimental environment.

In LTspice, set the global temperature to a certain value, which is also assumed to be the junction temperature. Next, set the output of the current source I_S and drive IGBT discrete with a single pulse. Then measure the on-state voltage under this condition. Afterward, tune the collector current I_C and junction temperature T_j (global temperature) to obtain the on-state voltages V_{CEon} under different conditions. In the simulation, this task can be completed quickly by using the sweep function. The value range of collector current I_C is from 5 A to 75 A, and junction temperature T_j is from 25°C to 175°C. The sample intervals are one ampere (1 A) and one-degree centigrade (1°C), respectively. Hence, 10721 groups of data in total can be obtained. One thousand groups for the training set and 20 groups for the test set are randomly assigned, respectively. Random sampling intends to check the neural network's robustness, but it is recommended to sample evenly in the real experiment.

3. Methods

This section introduces three methods, including backpropagation (BP) neural network, interpolation, and the Kalman Filter. The BP estimation and interpolated value are used as the Kalman filter measurement and prediction model, respectively. The setup and application of each method are also drawn.

3.1. Backpropagation Neural Network. Backpropagation (BP) neural network is a multilayer feedforward neural network trained according to the error backpropagation algorithm. It is capable of classifying arbitrary complex patterns and mapping multidimensional functions. The BP neural network base is the gradient descent method, which uses gradient search to minimize the mean square error between the actual and the expected output. The inputs are transmitted from the input layer to the output layer after being processed by the hidden layers during forward propagation. If the actual output is inconsistent with the expected output, then move to backpropagation. During backpropagation, the output is back transmitted to the input layer somehow, and the error is distributed to all units of each layer. Then, each layer's error can be obtained, which is used to correct the weight of each unit. With the continuous correction of the error, the network's accuracy will be improved step by step [20].

The junction temperature prediction model can be regarded as a complex nonlinear system, which is difficult to be accurately modelled with a single mathematical method. In this case, BP neural network can be constructed to express it. On-state voltage V_{CEon} and collector current I_C of IGBT discrete are chosen as the inputs. The junction temperature T_j is chosen to be the output. When the signal is transmitted, the inputs V_{CEon} and I_C act on the output node through the hidden layer. After the nonlinear transformation, the prediction of junction temperature T_e is generated from the output layer. If the actual output T_e is equal to the expected

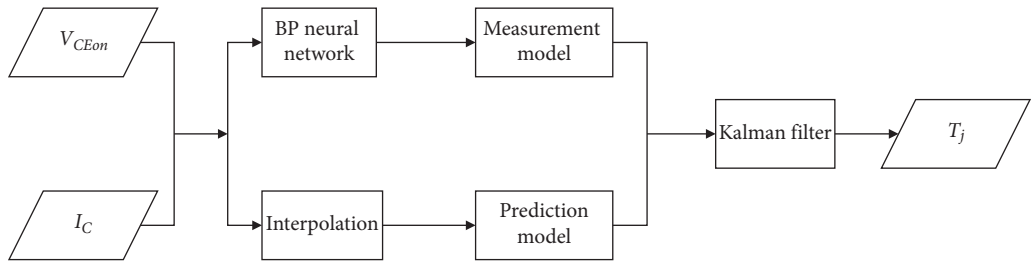


Figure 1: Block diagram of the methods for predicting junction temperature.

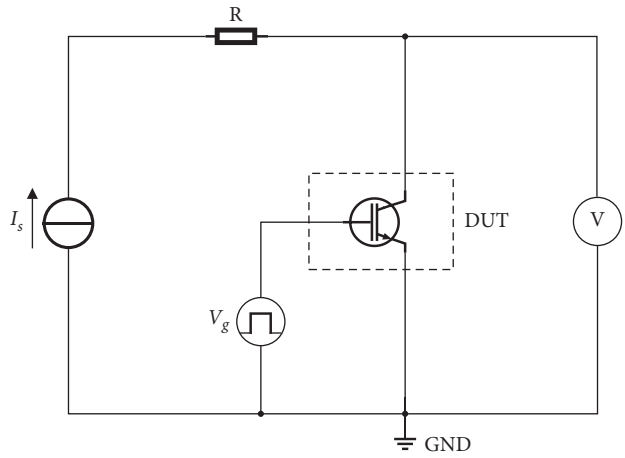


Figure 2: The simulation circuit for measuring on-state voltage.

output T_j, then propagation is terminated. Otherwise, the error will be allocated to all nodes in each layer through backpropagation. The neural network's weight and deviation are updated in the fastest increasing direction to minimize the error.

The structure of the BP neural network in this article is shown in Figure 3. As the figure shows, the neural network consists of two inputs, one output, and one hidden layer including ten neurons. For a complex task, choose 0.01 for learning rate and 0.9 for the momentum parameter to achieve high performance [20]. The other parameter settings for training the neural network are listed in Table 1, and the implementation is based on the MATLAB "neural network training toolbox." It is worth mentioning that, before starting the training, the dataset should be normalized to avoid possible numerical problems [21]. The BP estimation will be used in Section 3.3 as the measurement model of the Kalman filter.

3.2. Interpolation Method. The interpolation method interpolates the continuous function based on the discrete data. The continuous curve passes through all data points, which is a vital approach to approximate discrete functions. It can estimate the approximation of other points by analyzing the function value of finite points.

Use the interpolation function in MATLAB to generate an interpolation graph composed of on-state voltage V_{CEon}, collector current I_C, and junction temperature T_j, as shown in Figure 4. Because of the resolution limitation, the extracted temperature will be slightly different from the real junction temperature. On top of that, the personal error also probably exists during the extraction. The interpolated value will be used in Section 3.3 as the prediction model of the Kalman filter.

3.3. Kalman Filter. Kalman filter (KF) is an algorithm using the linear system's state equation to predict the system state through measurement. Since the measurement includes some noise and disturbance, the optimal estimation can also be regarded as a filtering process [22]. The first step is to predict the current state based on the previous state and control vector. The state equation from time $k-1$ to k is defined as

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1}, \tag{1}$$

where A and B are the state transition matrix and control-input matrix, respectively, A is set as identity matrix I, and control vector u_{k-1} is set as zero since the temperature does not change in a short time. The w_{k-1} is the noise during the prediction process. It is assumed to be white Gaussian noise with mean zero and covariance Q, denoted as $w_{k-1} \sim N(0, Q)$. Because of the limited resolution and personal error, the interpolated value will have some fluctuations. Hence, it can be considered to satisfy the Gaussian distribution with covariance Q.

The measurement equation defines the relationship between the state and the measurement at the time k as follows:

$$z_k = Hx_k + v_k, \tag{2}$$

where v_k is the noise from the measurement. It is assumed to be white Gaussian noise with mean zero and covariance R, denoted as $v_k \sim N(0, R)$. The measurement will be directly loaded from the BP estimations. The neural network's estimations can also be considered to satisfy the Gaussian distribution with covariance R since it has the oscillation around a fixed point.

When both the prediction and measurement model satisfy Gaussian distribution, their product will also be Gaussian distribution. The fused Gaussian distribution has a higher probability density and smaller variance, as shown in Figure 5. Kalman algorithm is a recursive prediction-update method and can be divided into prediction stage and correction stage. The prediction stage calculates the state variable's prior estimate based on the posterior estimate of the

previous moment. It can be described in the next two equations:

$$\hat{x}_{\bar{k}} = A\hat{x}_{k-1} + Bu_{k-1}, \tag{3}$$

$$P_{\bar{k}} = AP_{k-1}A^T + Q, \tag{4}$$

where P is the state error covariance; it represents the credibility of the prediction stage.

The correction stage combines the prior estimate with the new measurement variables to construct the optimal estimate. It can be described in the following three equations:

$$K_k = \frac{P_{\bar{k}}H^T}{HP_{\bar{k}}H^T + R}, \tag{5}$$

$$\hat{x}_k = \hat{x}_{\bar{k}} + K_k\left(z_k - H\hat{x}_{\bar{k}}\right), \tag{6}$$

$$P_k = \left(I - K_kH\right)P_{\bar{k}}, \tag{7}$$

where R is the covariance of the measurement noise; it represents the measurement stage's credibility.

The implementation of the Kalman filter is based on the MATLAB code. Use the interpolated value as the prediction and 100 different BP estimations to measure the Kalman filter, respectively. R and Q's values need to be tuned to obtain the most appropriate result, that is, make the optimal estimation closer to the expected value.

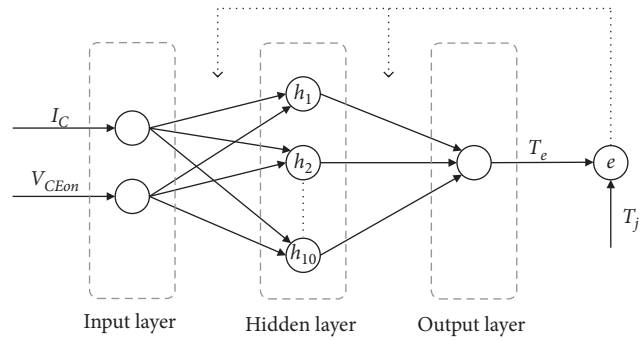


Figure 3: The flowchart of the BP neural network for estimating junction temperature.

Table 1: BP neural network training parameters.

Items	Parameters
Nodes in input layer	2
Nodes in hidden layer	10
Nodes in output layer	1
Activation function	Tansig
Output function	Purelin
Training function	Levenberg–Marquardt
Learning rate	0.01
Maximum epochs	1000
Performance goal	10^{-5}
Validation failures	6
Momentum factor	0.9

4. Results and Discussion

This section provides the validities of the BP neural network and the Kalman filter. Also, the detailed results of the three methods introduced in Section 3 are given. In the end, the experimental conclusions are also drawn.

4.1. Inflection Point of IGBT. Based on the simulation's collected data, the relationship between on-state voltage V_{CEon}, collector current I_C, and junction temperature T_j is drawn. Figure 6 shows that these coordinate points can make up a smooth surface. It is apparent that the value of V_{CEon} is affected by both I_C and T_j, which verifies the correctness of selecting V_{CEon} and I_C as the inputs of the temperature prediction model.

Figure 7 indicates that on-state voltage increases with the increment of junction temperature in the above three conditions but decreases below three conditions. This normal phenomenon is caused by the manufacturing process, which is known as the inflection point. The curves with positive and negative temperature coefficients intersect at this point.

By checking the data table obtained from simulation, the inflection point is found around 42 A. On-state voltage and junction temperature have a positive correlation when the collector current is larger than 42 A and a negative correlation when the collector current is smaller than 42 A. Consequently, the next sections' analysis will be divided into two parts ($I_C < 42$ A and $I_C > 42$ A). Limited to the article's length, this article's analysis will focus on the condition when the collector current is larger than 42 A. Still, the result of both conditions will be given in Section 4.5.

4.2. Validity of BP Neural Network. Sometimes, the BP neural network can only accurately predict specific but not all test sets because of contingency. To avoid that, it is necessary to pick different training and test sets to check the validity [23]. In this case, three completely different training and test sets are used to construct and evaluate the neural network. Each training set contains 1000 groups of data, and each test set includes 20 data groups. Besides, the parameter settings of the three control groups are the same.

The three control groups' error and percentage error are shown in Figures 8(a) and 8(b). It is observed that the absolute errors are mostly below 10°C and the percentage errors are mostly below 15%. The neural network performs well with each dataset, which also confirms its strong generalization ability.

4.3. Oscillation of BP Neural Network Estimations. Because the BP neural network model is initialized when built, the estimated value will differ each time. Pick three test data and put them in 100 different neural networks. The results are shown in Figure 9, and three test data are coloured differently. It is observed that the BP estimations are unstable but oscillate around their mean values. The oscillation appears because the initial weights and thresholds are

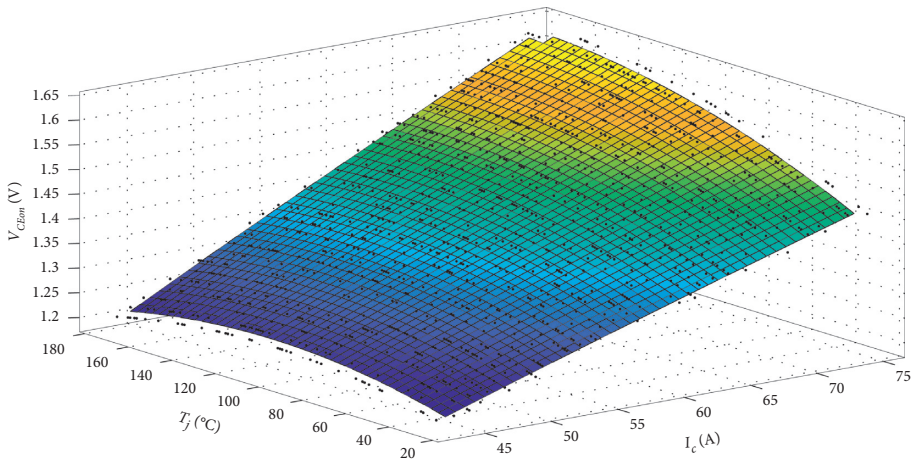


Figure 4: The interpolation graph which is composed of the training set.

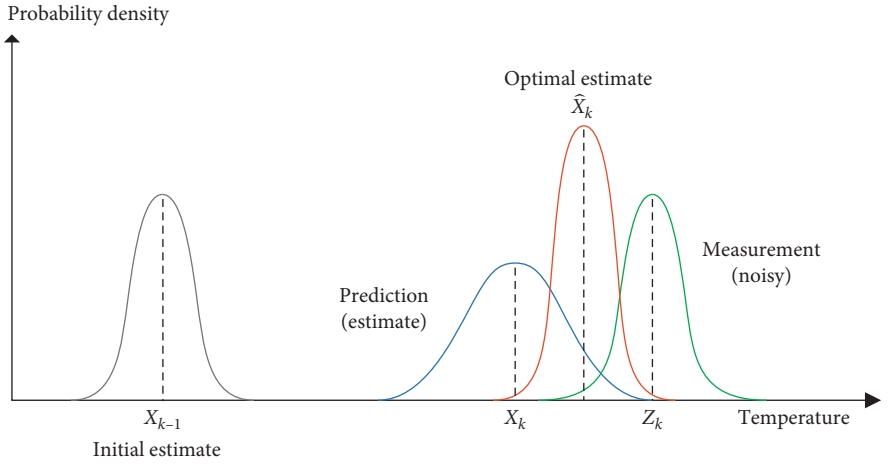


Figure 5: The working principle of Kalman filter from time $k-1$ to time k.

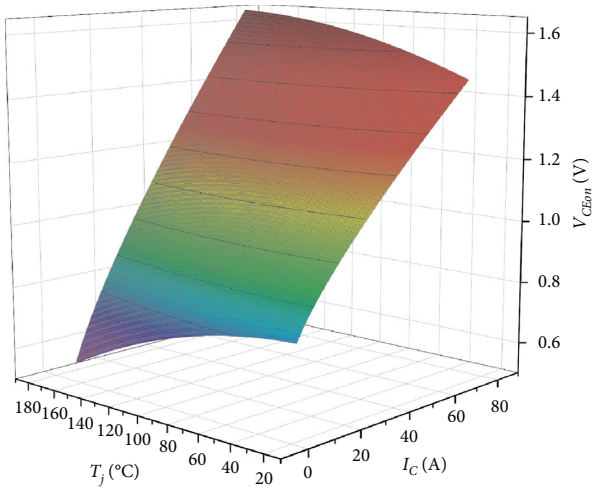


Figure 6: The three-dimensional space which is composed of junction temperature, collector current, and on-state voltage.

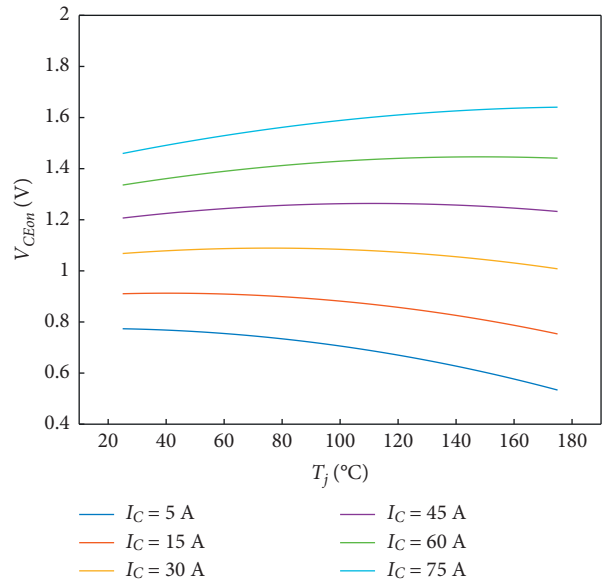


Figure 7: The relationship between junction temperature and on-state voltage with different collector currents.

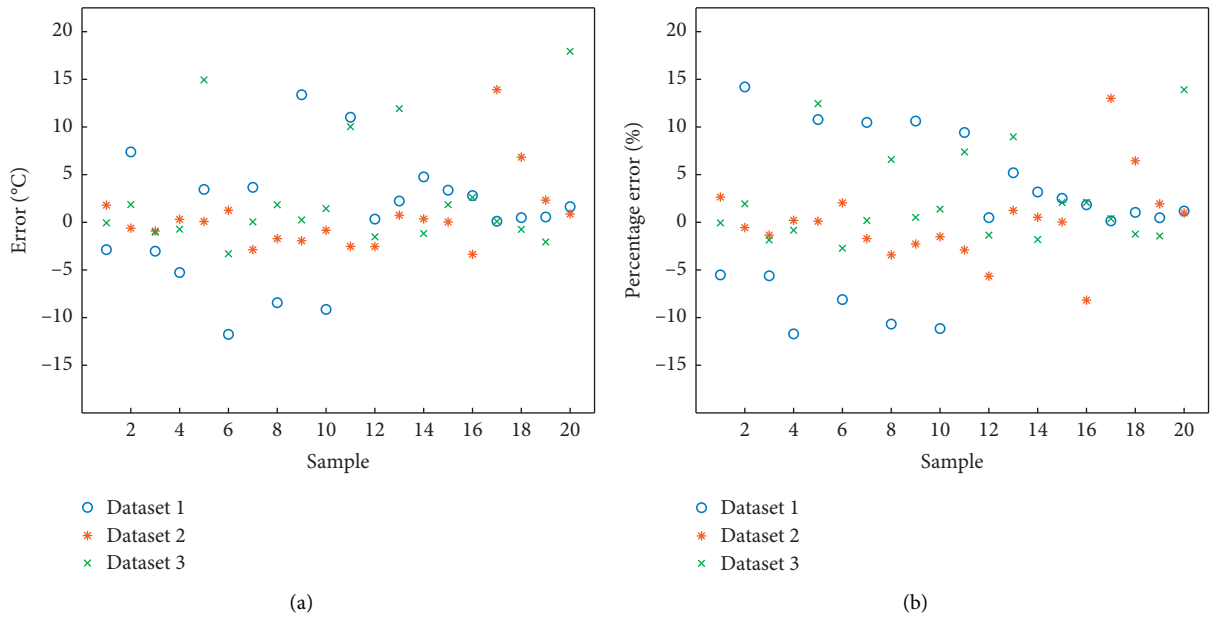


Figure 8: (a) The temperature errors of 20 samples. (b) The percentage errors of 20 samples.

generated randomly. Most of the points are close, but a few are far from the mean value. In other words, the closer to the mean, the higher probability of a point occurs. It indicates that the BP estimations used as the Kalman filter measurement model conform to the Gaussian distribution approximately. It is also found that the mean value of estimations becomes stable after around 15 repetitions. Thus, using the mean value is a feasible way to stabilize the oscillation. However, the estimation given by BP neural network can be further treated to increase the accuracy.

4.4. Validity of Kalman Filter. The BP estimations obtained from Section 4.3 are used as the measurement of the Kalman filter. The interpolated value obtained from Section 3.2 is used as the initial estimate of the Kalman filter. Pick one test data randomly to check the performance of the Kalman filter based on BP and interpolation. The results are shown in Figure 10.

The figure shows the performance comparison between interpolation, BP neural network, and Kalman filter. It is observed that the curve of Kalman filter prediction moves

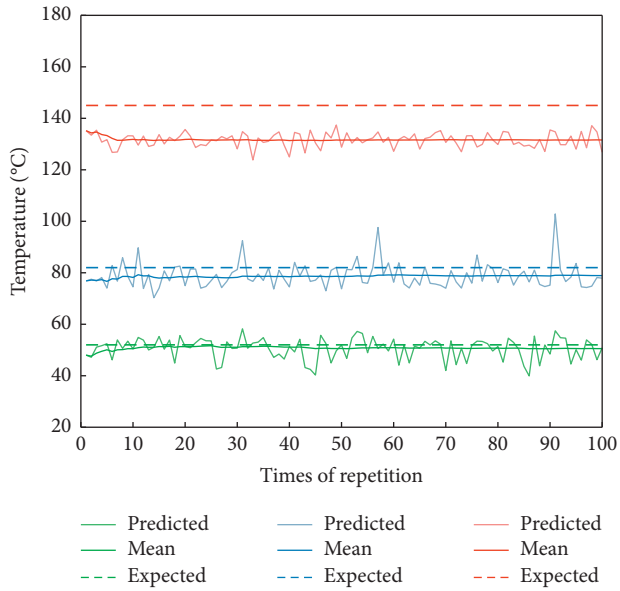


Figure 9: The temperature estimations of three samples with 100 times repetition.

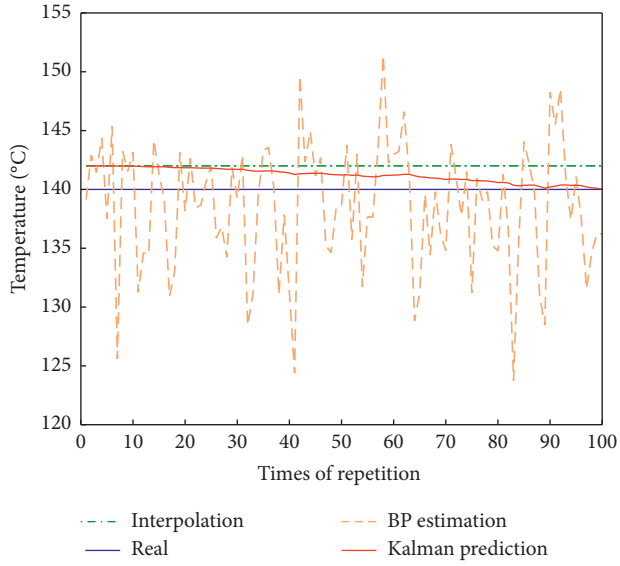


Figure 10: Performance comparison with 100 times of repetition.

towards where the BP estimations occur more frequently. Also, it converges after about 85 iterations and will be finally stabilized around the expected value. In this case, Kalman filter prediction has a big advantage over the other two methods in both stability and accuracy.

4.5. Performance Comparison between Three Predicting Methods. Check the Kalman filter's performance on all test data and compare the result with the mean value of BP estimations and the interpolated value. As shown in Figures 11(a) and 11(b), the Kalman filter greatly stabilizes the BP estimations and interpolated values. In both conditions ($I_C < 42$ A and $I_C > 42$ A), the errors are mostly below 5°C. The detailed predicted values in the two conditions are given in Tables 2 and 3. In most cases, the Kalman filter prediction is between the BP mean and interpolated value because the nature of the Kalman filter is the weighted average.

The comparison of errors is shown in Table 4. In condition $I_C < 42$ A, RMSE and MAPE of Kalman filter prediction are 2.6415 and 0.0166, respectively, which are smaller than the other two predicting methods. In condition $I_C > 42$ A, RMSE and MAPE of Kalman filter prediction are 4.8282 and 0.0284, respectively, which are also smaller than the other two predicting methods. The results indicate that the Kalman filter has a significant advantage in predicting junction temperature. The feasibility of using the Kalman filter based on BP neural network and interpolation has been further confirmed.

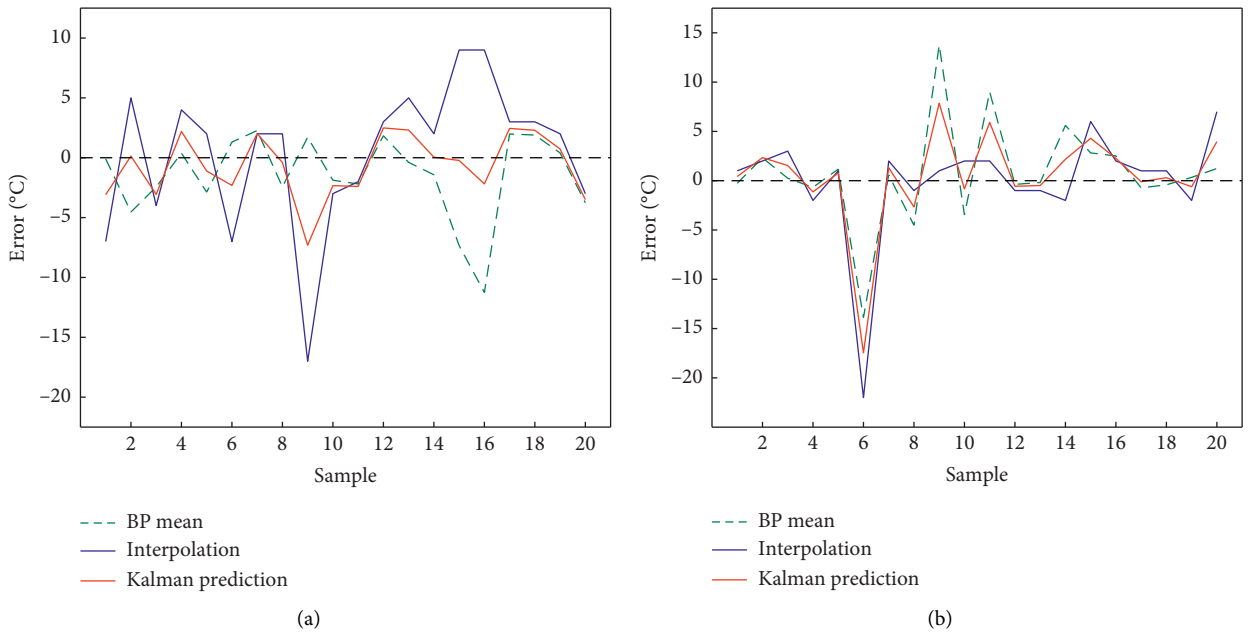


Figure 11: (a) The comparison of errors when $I_C < 42$ A. (b) The comparison of errors when $I_C > 42$ A.

Table 2: The numerical comparison of three methods when $I_C < 42$ A.

Test data	Real	BP mean	Interpolation	Kalman
1	173	172.9056	166	169.9133
2	99	94.4629	104	99.1401
3	174	171.5643	170	170.9382
4	102	102.3818	106	104.1974
5	142	139.1298	144	140.889

Table 3: The numerical comparison of three methods when $I_C > 42$ A.

Test data	Real	BP mean	Interpolation	Kalman
1	153	155.5118	155	155.267
2	76	75.2725	77	75.8904
3	47	46.5827	48	47.303
4	117	117.3351	115	116.39
5	139	140.2383	146	142.9634

5. Discussion

In Section 4.2, the validity of the BP neural network has been confirmed. The absolute errors of estimation are mostly less than 10°C, but the oscillation still impedes the accurate prediction of junction temperature. Calculating their mean can reduce the error and stabilize the oscillation to some extent. As Table 4 shows, the RMSE and MAPE of BP estimation are 3.6528 and 0.0221 ($I_C < 42$ A), respectively. The RMSE and MAPE of the interpolated value are 5.9161 and 0.0349 ($I_C < 42$ A), respectively. However, the BP neural network estimation and interpolation accuracy will be reduced in the real experiment because it is hard to obtain over 1000 samples. Using the Kalman filter based on BP neural network and interpolation can further stabilize the oscillation and reduce the error below 5°C. The RMSE and MAPE of Kalman filter prediction are 2.6415 and 0.0166 ($I_C < 42$ A), respectively. As one can see, Kalman filter prediction performs better than interpolation or BP estimation. In another condition ($I_C > 42$ A), the three predicting methods perform similarly.

Table 4: RMSE and MAPE comparisons of three predicting methods.

Method	Condition	RMSE[1]	MAPE[2]
BP mean	$I_C < 42$ A	3.6528	0.0221
	$I_C > 42$ A	5.2501	0.0309
Interpolation	$I_C < 42$ A	5.9161	0.0349
	$I_C > 42$ A	5.5588	0.0335
Kalman filter	$I_C < 42$ A	2.6415	0.0166
	$I_C > 42$ A	4.8282	0.0284

[1]Root mean square error. [2]Mean absolute percentage error.

The prediction of junction temperature is important for condition monitoring and degradation of the IGBT devices. Section 4.2 indicates it is feasible to estimate the junction temperature using BP neural network without modifying the standard package. Section 4.5 shows that the Kalman filter prediction accuracy in both conditions ($I_C < 42$ A and $I_C > 42$ A) is higher than the BP estimation or interpolation. In addition to increasing accuracy, it also enhances robustness. Even when outliers appear, the prediction filtered by the Kalman filter can still maintain stability.

Furthermore, there are several possibilities for improvement in this scheme. First, the updated version of the BP neural network such as PSO-BP [24], GA-BP [25], and MEA-BP [26] can be applied to increase the speed and accuracy of the convergence. Second, the use of the combined TSEPs can increase the reliability of the prediction [27, 28]. Third, other prediction models can be used instead

of interpolation, because the accuracy of interpolation could be greatly affected by the sample size.

In sum, both BP neural network and Kalman filter can work well in predicting IGBT junction temperature. The Kalman filter method further enhances accuracy and robustness. Nevertheless, compared with BP neural network, the Kalman filter based on it requires more complex processes. With the rapid development of deep learning or neural network, the accuracy of junction temperature prediction is expected to be further improved. Theoretically, the neural network can approach any complex function perfectly. What is more, the application scope of deep learning can be expanded to evaluate the aging degree or failure rate of the power device.

6. Conclusions

The Kalman filter based on BP neural network and interpolation proposed in this article has the following advantages:

(1) There is no need to modify the standard module package.

(2) It is simpler than the conventional TSEP method. The voltage drop between the junction and the measurement point can be neglected.

(3) It is more accurate and stable than the BP neural network estimation.

(4) It can be migrated to online monitoring after the entire prediction model has been built.

(5) There is a large room for improvement.

Deep learning in power electronics devices can help monitor the operating condition and evaluate the degradation from a new perspective. It is expected to promote the development of power electronics further.

Data Availability

The data and MATLAB code used to support this research article are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. Stippich, "Key components of modular propulsion systems for next generation electric vehicles," *CPSS Transactions on Power Electronics and Applications*, vol. 2, no. 4, pp. 249–258, 2017.

[2] A. Asadi and F. Pourfattah, "Effects of constructal theory on thermal management of a power electronic system," *Science Reports*, vol. 10, p. 21436, 2020.

[3] Electronics Prognostics 2020, https://ti.arc.nasa.gov/tech/dash/groups/pcoe/electronics-prognostics/.

[4] H. Wang, M. Liserre, and F. Blaabjerg, "Toward reliable power electronics: challenges, design tools, and opportunities," *IEEE Industrial Electronics Magazine*, vol. 7, no. 2, pp. 17–26, 2013.

[5] M. H. M. Sathik and S. Prasanth, "Comparison of IGBT junction temperature measurement and estimation methods-a review," *Asian Conference on Energy, Power and Transportation Electrification (ACEPT)*, vol. 1–8, 2017.

[6] Y. Avenas, L. Dupont, and Z. Khatir, "Temperature measurement of power semiconductor devices by thermo-sensitive electrical parameters-a review," *IEEE Transactions on Power Electronics*, vol. 27, no. 6, pp. 3081–3092, 2012.

[7] L. Dupont, Y. Avenas, and P.-O. Jeannin, "Comparison of junction temperature evaluations in a power IGBT module using an IR camera and three thermosensitive electrical parameters," *IEEE Transactions on Industry Applications*, vol. 49, no. 4, pp. 1599–1608, 2013.

[8] J. Zarebski and K. Gorecki, "The electrothermal large-signal model of power MOS transistors for SPICE," *IEEE Transactions on Power Electronics*, vol. 25, no. 5, pp. 1265–1274, 2010.

[9] L. M. Hillkirk, "Dynamic surface temperature measurements in SiC epitaxial power diodes performed under single-pulse self-heating conditions," *Solid-State Electronics*, vol. 48, no. 12, pp. 2181–2189, 2004.

[10] X. Perpiñà, J. F. Serviere, J. Saiz, D. Barlini, M. Mermet-Guyennet, and J. Millán, "Temperature measurement on series resistance and devices in power packs based on on-state voltage drop monitoring at high current," *Microelectronics Reliability*, vol. 46, no. 9–11, pp. 1834–1839, 2006.

[11] W. Brekel, "Time resolved in situ Tvj measurements of 6.5 kV IGBTs during inverter operation," in *Proceedings PCIM Europe*, pp. 808–813, Nuremberg, Germany, May 2009.

[12] Z. Hu, W. Zhang, and J. Wu, "An improved electro-thermal model to estimate the junction temperature of IGBT module," *Electronics*, vol. 8, no. 10, p. 1066, 2019.

[13] I. Bahun, V. Sunde, and Z. Jakopovic, "Estimation of insulated-gate bipolar transistor operating temperature: simulation and experiment," *Journal of Power Electronics*, vol. 13, no. 4, pp. 729–736, 2013.

[14] U. M. Choi, F. Blaabjerg, F. Iannuzzo, and S. Jørgensen, "Junction temperature estimation method for a 600 V, 30A IGBT module during converter operation," *Microelectronics Reliability*, vol. 55, no. 9-10, pp. 2022–2026, 2015.

[15] Z. Xu, F. Xu, and F. Wang, "Junction temperature measurement of IGBTs using short circuit current as a temperature sensitive electrical parameter for converter prototype evaluation," *IEEE Trans. Ind. Electron.*vol. 62, pp. 3419–3429, 2014.

[16] N. Baker, S. Munk-Nielsen, F. Iannuzzo, and M. Liserre, "IGBT junction temperature measurement via peak gate current," *IEEE Transactions on Power Electronics*, vol. 31, no. 5, pp. 3784–3793, 2016.

[17] A. Amoiridis, A. Anurag, P. Ghimire, S. Munk-Nielsen, and N. Baker, "Vce-based chip temperature methods for high power IGBT modules during power cycling—a comparison," in *Proceedings of the 2015 17th European Conference on Power Electronics and Applications*, pp. 1–9, EPE'15 ECCE-Europe), Geneva, Switzerland, September 2015.

[18] C. Dong and P. Mao, "The junction temperature measurement of insulated gate bipolar transistor based on multi-layer feed -forward neural network is presented," in *Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 136–140, Beijing, China, August 2020.

[19] P. M. Granitto, P. F. Verdes, and H. A. Ceccatto, "Neural network ensembles: evaluation of aggregation algorithms," *Artificial Intelligence*, vol. 163, no. 2, pp. 139–162, 2005.

[20] A. Suliman and Y. Zhang, "A review on back-propagation neural networks in the application of remote sensing image classification," *Journal of Earth Science and Engineering*, vol. 5, pp. 52–65, 2015.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, Sardinia, Italy, May 2010.

[22] C. Campestrini, T. Heil, S. Kosch, and A. Jossen, "A comparative study and review of different Kalman filters by applying an enhanced validation method," *Journal of Energy Storage*, vol. 8, pp. 142–159, 2016.

[23] Y. Duan and X. Chen, "Benchmarking deep reinforcement learning for continuous control," 2016, https://arxiv.org/abs/1604.06778.

[24] J. Huang and L. He, "Application of improved PSO - BP neural network in customer churn warning," *Procedia Computer Science*, vol. 131, pp. 1238–1246, 2018.

[25] Y. H. Yang, L. Peng, X. B. Chen, and X. Z. Liu, "A GA-BP neural network model for predicting the temperature of slabs in the reheating furnace," *Applied Mechanics and Materials*, vol. 58-60, pp. 1371–1377, 2011.

[26] J. L. Gai, "Modeling of photovoltaic cell based on BP neural networks improved by MEA," *Applied Mechanics and Materials*, vol. 219, pp. 809–814, 2012.

[27] L. Shao, Y. Hu, and G. Xu, "A high precision on-line detection method for IGBT junction temperature based on stepwise regression algorithm," *IEEE Access*, vol. 8, pp. 186172–186180, 2020.

[28] D. Herwig, T. Brockhage, and A. Mertens, "Combining multiple temperature-sensitive electrical parameters using artificial neural networks," in *Proceedings of the 2020 22nd European Conference on Power Electronics and Applications (EPE'20 ECCE Europe)*, Paris, France, September 2020.

Hindawi
Complexity
Volume 2021, Article ID 6696064, 9 pages
https://doi.org/10.1155/2021/6696064

Research Article

MAF-CNER : A Chinese Named Entity Recognition Model Based on Multifeature Adaptive Fusion

Xuming Han,[1] **Feng Zhou**,[2] **Zhiyuan Hao**,[3] **Qiaoming Liu**,[4] **Yong Li**,[2] **and Qi Qin**[2]

[1]*College of Information Science and Technology, Jinan University, Guangzhou 510632, China*
[2]*School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China*
[3]*School of Management, Jilin University, Changchun 130022, Jilin, China*
[4]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China*

Correspondence should be addressed to Zhiyuan Hao; 15391910163@163.com and Qiaoming Liu; cslqm@hit.edu.cn

Received 18 December 2020; Revised 27 December 2020; Accepted 11 January 2021; Published 9 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

Named entity recognition (NER) is a subtask in natural language processing, and its accuracy greatly affects the effectiveness of downstream tasks. Aiming at the problem of insufficient expression of potential Chinese features in named entity recognition tasks, this paper proposes a multifeature adaptive fusion Chinese named entity recognition (MAF-CNER) model. The model uses bidirectional long short-term memory (BiLSTM) neural network to extract stroke and radical features and adopts a weighted concatenation method to fuse two sets of features adaptively. This method can better integrate the two sets of features, thereby improving the model entity recognition ability. In order to fully test the entity recognition performance of this model, we compared the basic model and other mainstream models on Microsoft Research Asia (MSRA) and "China People's Daily" dataset from January to June 1998. Experimental results show that this model is better than other models, with F1 values of 97.01% and 96.78%, respectively.

1. Introduction

Word representation learning has been widely concerned as a basic problem in the field of natural language processing. Unlike traditional one-hot representations, low-dimensional distributed vocabulary representations (also called word embeddings) represent words as low-dimensional dense real number vectors, which can better capture the associated information between natural language words. This form of representation is very useful in some downstream tasks of natural language processing, for example, text classification [1], NER [2, 3], relation extraction [4, 5], and sentiment analysis [6, 7]. Therefore, how to obtain a better semantic representation of words is crucial.

In recent years of research, the NER model is mainly based on deep learning, and with the development of deep learning, more and more remarkable results have been achieved. The main basic model framework of English NER is Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) [8], which uses word embedding as the basic unit of predicting labels. English is a kind of phonetic alphabet, but Chinese characters represent typical meanings, so these research methods cannot be directly applied to Chinese. Unlike English, Chinese sentences do not have as obvious separators as in English. Therefore, when processing Chinese NER tasks, first use word segmentation tools to segment sentences, and then implement a sequence tagging model based on segmentation words. This method results in poor performance of CNER because CNER faces the following difficulties: (1) the quality of sentence segmentation has a great impact on the performance of NER. For example, "武汉市长江大桥" (Wuhan Yangtze River Bridge) as a whole location named entity, after segmentation by word segmentation tools, it may be segmented into "武汉" (Wuhan), "市长" (mayor), and "江大桥" (river's bridge). When these participles are used as input to the NER model, they will be recognized as three different named entities. (2) In order to solve the problem of word-

level embedding, character-level embedding is widely used in NER tasks, but it still has many shortcomings. As shown in Figure 1, "人" (people), "八" (eight), and "乂" (yi) are semantically unrelated, and the stroke sequence is "丿" (left-falling stroke) and "㇀" (right-falling stroke). Chinese characters with the same stroke sequence have completely different semantics, and the stroke sequence of a Chinese character cannot uniquely identify a Chinese character. Similarly, using the radical feature alone will also encounter the same problem. In order to solve this problem, we can introduce another internal characteristic of Chinese characters—the roots; the radicals of "人", "八", and "乂" are "人", "八" , and "丿" , respectively. Combining the two characteristics of this stroke and the radical can distinguish Chinese characters well. Figure 1

Integrating the internal characteristics of Chinese characters is effective for learning Chinese word embedding [9]. For example, Yin et al. used Convolutional Neural Networks (CNNs) to extract radical features, aiming to capture the intrinsic and intrinsic correlation of characters. Experimental results show that the model has achieved good performance in the field of Chinese clinical NER [10]. Chinese characters have rich internal structural features. How to better learn and use these features to improve the quality of Chinese character embedding is very important. It can be further studied on how to better combine the character characteristics of Chinese characters with the internal characteristics of Chinese characters. This article designs a multifeature adaptive fusion (MAF) method to fuse the stroke features and radical features of Chinese characters. This method can adaptively calculate the weight of the fusion stroke feature and radical feature. The main contributions of this article can be summarized as follows:

(1) This article integrates the characteristics of characters, strokes, and radicals into the BiLSTM-CRF model to fully represent the semantic information of Chinese characters.

(2) To achieve a better and more balanced fusion of the two sets of features, this article adopts the method of weighted series adaptive fusion features.

(3) Evaluation results show that this model achieves good performance on both the MSRA dataset and the 1998 China People's Daily dataset.

The remainder of this article is structured as follows. Section 2 introduces the related work of CNER. Section 3 gives a detailed description of the MAF-CNER model. Section 4 presents extensive experiments to verify the effectiveness of our proposal, and Section 5 summarizes this work.

2. Related Works

The traditional solution to the NER problem mainly includes three methods: rule-based method, statistics-based method, and dictionary-based method [11]. The method based on rules and dictionaries requires professional linguists to write rules by hand, requires a lot of time, and has poor portability

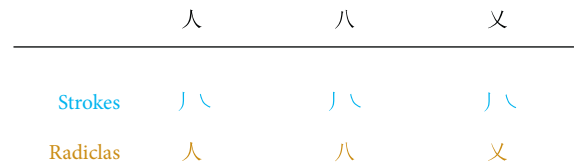


Figure 1: Strokes and radicals of "人", "八", and "乂".

in different fields. In the task of NER, statistical methods mainly use Conditional Random Field (CRF) and Hidden Markov Model (HMM) [12, 13]. Although the accuracy rate is improved compared with the method based on rules and dictionaries, it still has disadvantages such as long training time.

With the continuous development of deep learning, researchers began to apply deep learning to NER tasks. Compared with traditional models, neural network models can learn deeper semantic feature information with almost no need for feature engineering [14] and domain knowledge. These models further improve the accuracy of entity recognition, especially the BiLSTM-CRF model [15, 16], and can significantly improve the performance of NER tasks.

The standard model for solving NER problems in the English domain is the BiLSTM-CRF model proposed by Huang et al. [17], which is more robust and less dependent on word embedding. Based on this structure, Lample et al. proposed to use BiLSTM to extract word representations on character-level embedding. Cho et al. proposed a deep learning NER model that effectively represents biomedical word tokens through the design of a combinatorial feature embedding, enhanced by integrating two different character-level representations extracted from CNN and BiLSTM [18]. In the Chinese field, CNER is more challenging [19]. Wang et al. proposed a CNN model based on a gating mechanism (GCNN) [20]. Cao et al. used Chinese character strokes as features and proposed the stroke n-gram model, which not only excavated the feature information of Chinese character strokes but also more effectively used the semantic information of Chinese characters to train word vectors [21]. Cao et al. proposed a novel adversarial transfer learning framework to make full use of the boundary information shared by tasks and prevent the task-specific functions of Chinese word segmentation [22]. Xu et al. proposed a simple and effective neural network framework ME-CNER (Multiple Embeddings for Chinese Named Entity Recognition), which embeds rich semantic information at multiple levels from radicals, characters to words [23]. Wu et al. proposed a radical-based CNER RCBC (R-CNN-BiLSTM-CRF). The RCBC-based model uses CNNs to automatically extract the semantics of the radicals of Chinese characters and combines the word vectors and radical vectors into joint vector. This method can reduce the semantic deviation of radical features and capture semantic information more accurately [24]. Ye et al. proposed a CNER model based on character-word vector fusion. This model reduces the dependence on the accuracy of word segmentation algorithms and effectively utilizes the semantic features of words [25]. In order to solve the ambiguity of Chinese words and the lack of word boundaries, Wu et al. proposed a novel fine-grained

character-level representation method to capture the semantic information of Chinese characters [26]. Although the above methods have achieved good results, none of them have a more in-depth exploration of the internal characteristics of Chinese characters, and the fusion methods between multiple characteristics can be studied more deeply.

3. MAF-CNER Model

This section introduces the network layer organization structure of the "multifeature adaptive fusion Chinese named entity recognition model" model, as shown in Figure 2. The model is divided into three layers: character, stroke, and radical multifeature vector fusion layer; BiLSTM layer; CRF layer. The radical and stroke feature representations are calculated by the BiLSTM neural network, merged using the weighted concatenation method and concatenated with the character vector to form the final input vector. BiLSTM extracts the context features of the current input vector. The input of the CRF layer is the output vector of the BiLSTM layer, and the CRF layer will decode the information and obtain the best tag sequence. We will introduce the components of the Chinese clinial NER model based on MAF from bottom to top, as follows.

3.1. Character, Stroke, and Radical Multifeature Vector Fusion Layer. For a given sentence sequence $x = (c_1, c_2, \ldots, c_n)$, the embedding vector is composed of Chinese characters. Character characteristics are $e_c \in R^{d_c}$, radical characteristics, $e_r \in R^{d_r}$, and stroke characteristics, $e_s \in R^{d_s}$, respectively. As shown in Figure 3, the embedding vector of each character C_i can be expressed as follows:

$$e_i = e_i^c \oplus \left(m * e_i^r \oplus n * e_i^s\right). \tag{1}$$

3.2. Character Embedding. Character-level embedding has been widely used in natural language processing. Research shows that embedding pretrained characters in a specific field can improve system performance. For example, adding character-level features in neural machine translation [27, 28] can improve the translation performance of the system, text classification [29, 30], and NER also uses character-level representation. Therefore, the pretrained character embedding is better than the random initial character embedding. This article uses the Chinese Wikipedia corpus of May 2020 to pretrain Chinese character embedding through Word2Vec. After preprocessing, about 171M training corpus is finally obtained. The pretraining of character embedding is implemented with the Python version of Word2Vec in Gensim, and the dimension of the feature vector is set to 100.

3.3. Adaptive Fusion Representation of Strokes and Radical Features

3.3.1. Radical Features. A Chinese character is a kind of pictograph, and the radical is the first stroke or shape of a Chinese character. One of the most notable

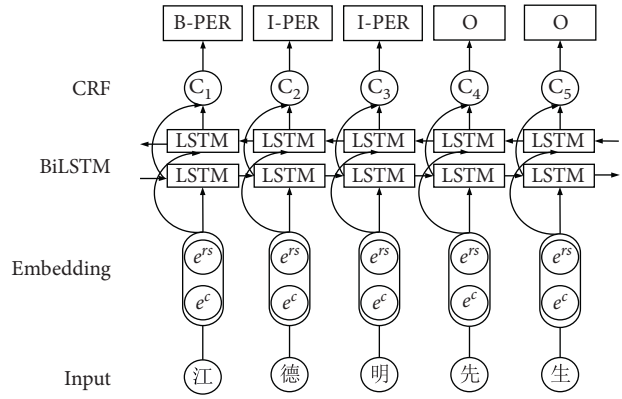


Figure 2: MAF-CNER model network structure.

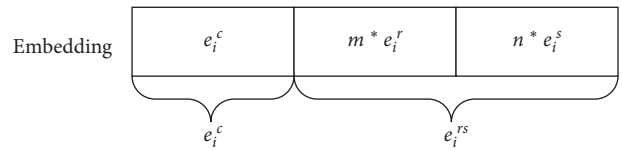


Figure 3: Characters, strokes, and radical vector links.

features of Chinese characters is that they contain a lot of semantic information at the radical level. The radicals of Chinese characters have a very important impact on the semantics of Chinese characters. For example, "胖" (fat) , "胸" (chest), and "肺" (lung). The main radical "月" (moon) is a simplified form of "肉" (flesh), which stands for meat, indicating that these characters are related to organs. A total of 228 radicals such as "鹿" (deer), "卤" (halogen), and "丶" (dot) are numbered from 1 to 228. However, the research in the traditional model mainly focuses on the semantic research at the phrase level.

This article uses the BiLSTM network to extract the semantic information of the corresponding radicals of Chinese characters. Figure 4 shows the overall structure of the model in detail. The expression is as follows:

$$e_r = \lfloor \overrightarrow{h_t}; \overleftarrow{h_t} \rfloor. \tag{2}$$

In formula 2, $\left\lfloor \overrightarrow{h_t}; \overleftarrow{h_t} \right\rfloor$ is the hidden layer vector obtained by training BiLSTM network.

3.3.2. Stroke Characteristics. Stroke usually refers to the uninterrupted dots and lines of various shapes that compose Chinese characters, such as horizontal ("一"), vertical ("丨"), and left-falling stroke ("丿") and dot ("丶"). It is the smallest continuous stroke unit that constitutes a Chinese character. As shown in Table 1, we divide the strokes into five types with the corresponding numbers of 1 to 5.

The Chinese character writing system provides a guide for the stroke order of each Chinese character. With this stroke information, we can decompose Chinese characters into strokes in a specific stroke order. This sequence information can be used when learning the internal semantic information of Chinese characters. Therefore, this article

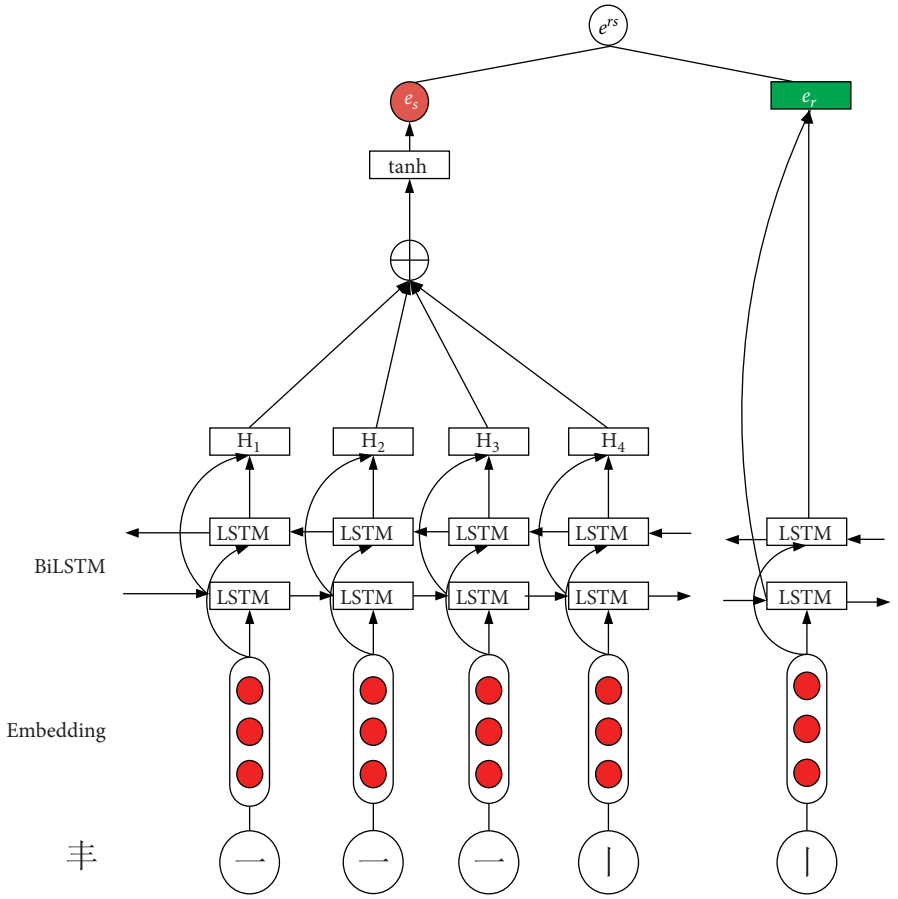


Figure 4: Stroke and radical feature calculation and fusion.

Table 1: Stroke type and number.

Stroke name	Horizontal	Vertical	Left-falling stroke	Right-falling stroke	Others
Type	一	丨	丿	丶	〈, 乛, 乚
Number	1	2	3	4	5

uses the BiLSTM network to extract the contextual semantic information of Chinese character strokes. Figure 4 shows the model structure. This method can learn more Chinese character graphic features. The expression is as follows:

$$e_s = \tanh\left(\sum_{j=1}^{N} H_{i,j}\right), \tag{3}$$

where $H_{i,j}$ is the j-th stroke feature vector of the i-th Chinese character.

3.3.3. Adaptive Feature Fusion. As shown in Figure 4, this article takes the stroke feature as the main feature, calculates its similarity with the character vector obtained by Word2Vec training, and determines its weight $\boldsymbol{m}$ according to formula 4.

$$m = e^{-\left(e_c \cdot e_s / \|e_c\| \times \|e_s\|\right)}, \tag{4}$$

where e_c is a character vector and e_s is a stroke vector.

The radical feature is used as an auxiliary feature, and the importance of the radical itself is calculated according to formulas 5 and (6), and its weight n is determined, and the weighted series method is used to fuse the two sets of features. This method can not only learn more graphic features of Chinese characters but also make the combination of the two features more balanced.

$$w_i = a \cdot e_r + b, \tag{5}$$

$$n = \text{sigmoid}(w_i). \tag{6}$$

In formula 5, a and $\boldsymbol{b}$ are trainable parameters.

The features of adaptive fusion are expressed as follows:

$$e_i^{rs} = me_i^s \oplus ne_i^r. \tag{7}$$

3.4. BiLSTM Layer. Long Short-Term Memory (LSTM) is a variant of Recurrent Neural Network (RNN), which has been widely used in many natural language processing (NLP) tasks, such as NER, text classification, and sentiment analysis. It introduces the cell state and uses input gates, forget gates, and output gates to maintain and control information, which can effectively overcome the gradient explosion and gradient loss caused by the long-distance dependence of the RNN model. The mathematical expression of the LSTM model is as follows:

$$\begin{aligned} i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i), \\ f_t &= \sigma\left(W_f[h_{t-1}, x_t] + b_f\right), \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o), \\ \widetilde{c_t} &= \tanh\left(W_c[h_{t-1}, x_t] + b_c\right), \\ c_t &= i_t * \widetilde{c_t} + f_t * c_{t-1}, \\ h_t &= o_t * \tanh(c_t), \end{aligned} \tag{8}$$

where σ represents the sigmod activation function. tanh represents the hyperbolic tangent function. X_t represents unit input. i_t, f_t, o_t represent the input gate, forget gate, and output gate at time W and b, respectively, which represent the weight and deviation of the input gate, forget gate, and output gate. $\widetilde{c_t}$ represents the current state of the input. c_t represents the update status at h_t that is the output at $\boldsymbol{t}$.

In order to use character context information at the same time, the model in this article uses BiLSTM to get the context vector of each character, which is a combination of forward LSTM and reverse LSTM. For a given sentence $x = (x_1, x_2, \ldots, x_n)$, we use $\overrightarrow{h_t}$ to represent the hidden layer state of the forward LSTM at time t, whereas h_t represents the reverse LSTM at time t . Hidden layer state, by linking the corresponding forward and reverse LSTM states, gets the final context vector $h_t = \left\lfloor \overrightarrow{h_t}; \overleftarrow{h_t} \right\rfloor$.

3.5. CRF Layer. Compared with the HMM, CRF does not have the strict requirements of the independence assumption of HMM and can effectively use both the internal information of the sequence and the external observation information, avoiding the problem of labeling bias and directly assuming the possibility of labeling and performing differentiated modeling. CRF can capture more dependencies: for example, "I-LOC" tags cannot follow "B-PER" [20]. In CNER, the input of CRF is the context feature vector learned from the BiLSTM layer. For input text sentence,

$$x = (x_1, x_2, x_3, \ldots, x_n). \tag{9}$$

Let $P_{i,j}$ denote the probability score of the j-th label of the i-th Chinese character in the sentence. For a prediction sequence $y = \{y_1, y_2, \ldots, y_n\}$, the CRF score can be defined as follows:

$$f(x, y) = \sum_{n=0}^{n+1} M_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}, \tag{10}$$

where M is the transition matrix and $M_{i,j}$ represents the transition score from label i to j. y_0 and y_{n+1} represent the start and end tags, respectively. Finally, we use softmax function to calculate the probability of the sequence y as follows:

$$P(y|x) = \frac{e^{f(x,y)}}{\sum_{\widetilde{y} \in Y_x} e^{f(x,y)}}. \tag{11}$$

During the training process, maximize the log probability of the correct label sequence:

$$\log(P(x, y)) = f(x, y) - \log\left(\sum_{\widetilde{y} \in Y_x} e^{f(x, \widetilde{y})}\right). \tag{12}$$

In the decoding stage, we predict that the maximum score obtained by the output sequence is as follows:

$$y^* = \underset{\widetilde{y} \in Y_x}{\operatorname{argmax}} f(x, \widetilde{y}). \tag{13}$$

In the prediction stage, the dynamic programming algorithm, Viterbi, is used to solve the optimal sequence.

4. Experiments and Results

4.1. Experimental Data and Evaluation Indicators. In order to evaluate the model proposed in this article on the task of CNER, this article conducted experiments on two different widely used datasets, namely, the MSRA dataset and the "China People's Daily" dataset from January to June 1998. Table 2 shows the statistical information of the data set used in this article.

4.1.1. MSRA. It is a general dataset for CNER. The dataset contains three named entities: PER (person), LOC (location), and ORG (organization). The training set contains 46364 sentences, and the test set contains 4365 sentences. This article uses the ternary tag set {B, I, O} to mark, B represents the first word of the entity, I represents the remaining words of the entity, and O represents the non-entity word.

4.1.2. China People's Daily. The China People's Daily corpus was released by the Institute of Computational Linguistics of Peking University from January to June 1998. The entity categories are PER (person), LOC (location), and ORG (organization), also using the ternary tag set {B, I, O} for labeling. This article uses the data from January to May 1998 as the training set and the validation set. The validation set is 1/5 of the total data from January to May. The data from June 1998 is used as the test set.

Table 2: Dataset statement summary table.

Data set	Number of sentences		
	Train	Dev	Test
MSRA	46364	—	4365
China People's Daily	82116	20529	21131

In order to fully evaluate the performance of the model, we use Precision (P), Recall (R), and harmonic average F1-score (F1) as the evaluation criteria for model performance, which is defined as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\%,$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\%, \tag{14}$$

$$F1 = \frac{2 * P * R}{P + R} \times 100\%.$$

TP (True Positive) indicates the correct number of samples in the positive examples, FP (False Positive) indicates the number of incorrect samples in the negative examples, and FN (False Negative) indicates the number of incorrect samples in the positive examples.

4.2. Model Building and Parameter Setting. The model in this article is built using PyTorch. PyTorch was launched by the Facebook Artificial Intelligence Research Institute (FAIR) in January 2017 based on Torch and is widely used in applications such as NLP. The experimental parameters are set as follows: embedding dimension (embedding_dim) is 300, input dimension max_length is 80, and training set batch_size: China People's Daily dataset is 100, MSRA dataset is 128, and MSRA dataset is 64. The training learning rate is set to 0.001, in order to prevent overfitting during training; the weight decay factor weight_decay is set to $5e^{-4}$, dropout technology is used to prevent overfitting, and the value is set to 0.5.

4.3. Experimental Results. In order to more objectively evaluate the model performance of this model on the MSRA dataset and the "China People's Daily" dataset, LSTM, BiLSTM, and BiLSTM-CRF models are used for performance analysis. The experimental results are shown in Table 3.

In Table 3, from the comparison of the experimental results of LSTM and BiLSTM, it can be seen that the latter performs better than the former. This verifies that the BiLSTM network can better capture the context information of the serialized text, with stronger learning ability that is better than LSTM . In the comparison between BiLSTM and BiLSTM-CRF, after adding the CRF module, it can be seen that the BiLSTM-CRF model has various aspects. Both are better than BiLSTM, which is mainly due to the fact that CRF considers the global label information in the sequence during the decoding process, which improves the performance of the model. Our model introduces two features of strokes and radicals on the basis of character-level embedding, and the test results on the two datasets achieve the best performance.

In order to verify the effectiveness of this method, it is compared with other mainstream NER methods. The specific results are shown in Tables 4 and 5. In Table 4, Chen et al. used CRF based on character features, and the F1 value was 86.20% [31]. The model of Zhou et al. used a multistage model. They used a character-level CRF model to segment the sequence. Then, word-level CRF layer was used to identify named entities, and the F1 value on the MSRA dataset reached 86.51% [32]. Zhou et al. took CNER as a joint recognition and classification task based on a global linear model [33]. The model used the rich manual feature model proposed in the literature [41] to greatly improve the performance of CNER. The F1 value of another BiLSTM-CRF neural network model proposed by Dong et al. was close to 90.95%. This model used both character-level and radical-level representations in the input of the model structure [34]. Zhang et al. used a lattice LSTM model for CNER. This model encodes the input character sequence and all possible words matching the dictionary. The F1 value of the model reached 93.18%, but the authors did not use the development dataset and trained the lattice LSTM mode [35]. Zhao et al. used a pretrained language model to encode the input sequence as a contextual representation and designed a new model that combines neural networks with BERT; the F1 value of the model reaches 95.28% [36]. However, using the model in this article, the F1 value reaches 97.01%. Johnson proposed the comprehensive embedding, which can take character, word, and position into account, has a valid structure, and can seize effective information. Regarding the test performance on MSRA dataset, F1 value reached 92.99%. Compared with the above model, this model has the best performance.

Table 5 shows test model performance using China People's Daily dataset. Collobert et al. used a feedforward neural network, combined with preprocessing, affixes, and capitalization features, and achieved a result of 88.50% F1 [38]; Lample et al. input character-level word vectors into the BiLSTM-CRF model and achieved F1 value of 90.08% [19]; Chiu et al. combined BiLSTM with the CNN model and

Table 3: Comparison results between the model in this article and the basic model.

MSRA				People's Daily		
Model	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
LSTM	84.56	83.25	83.90	87.25	85.76	86.49
BiLSTM	86.77	84.93	85.84	88.34	87.61	87.97
BiLSTM-CRF	89.45	87.20	88.31	90.45	89.72	90.08
Ours	**97.21**	**96.83**	**97.01**	**97.08**	**96.47**	**96.78**

Table 4: Experimental results of MSRA dataset.

Model	Precision (%)	Recall (%)	F1 (%)
Chen [31]	91.22	81.71	86.20
Zhou [32]	88.94	84.20	86.51
Zhou [33]	91.86	88.75	90.28
Dong [34]	91.28	90.62	90.95
Zhang [35]	93.57	92.79	93.18
Zhao [36]	95.46	95.09	95.28
Johnson [37]	93.71	92.29	92.99
Ours	**97.21**	**96.83**	**97.01**

Table 5: . Experimental results of China People's Daily.

Model	Precision (%)	Recall (%)	F1 (%)
Collobert [38]	88.43	87.68	88.05
Lample [19]	90.45	89.72	90.08
Chiu [39]	91.94	90.06	90.99
Shen [40]	91.46	90.18	90.81
Ours	**97.08**	**96.47**	**96.78**

achieved 91.49% of advanced results [39]. Shen et al. proved that when deep learning is combined with active learning, the amount of labeled training data can be reduced. Although active learning can improve sample efficiency, it may be computationally expensive due to iterative retraining. In order to speed up the introduction of a lightweight architecture, the CNN-CNN-LSTM model consists of a convolutional character and word encoder and a LSTM tag decoder [40]. This article uses BiLSTM-CRF as the basic model and introduces two kinds of internal semantic information of Chinese character strokes and radicals. Model performance F1 increased to 96.78%.

5. Conclusion

In view of the insufficient representation of potential features of Chinese characters, this article uses BiLSTM network to learn the internal strokes and radical semantic information of Chinese characters and combines with the BiLSTM-CRF model to construct an adaptive multifeature fusion embedded CNER model. The assessment was conducted on the MSRA corpus and the corpus of China People's Daily" from Janu"ary to June 1998. Compared with other mainstream methods, the model in this article achieves the best results on both corpora. The biggest advantage of this model is that the weighted concatenation method is used to adaptively fuse two kinds of semantic information in Chinese characters, while previous research only stayed at the word-level embedding or used one kind of internal characteristic semantic information of Chinese characters. This will make the embedding layer insufficiently represented, the performance of the model will be relatively reduced, and the named entity cannot be correctly identified. Combining the two internal features can make Chinese character features more fully represented, avoiding the problem where a single feature cannot correctly distinguish Chinese characters, and the proportion of the two semantic information combinations is more balanced through weighting, and the best combination effect is achieved.

Data Availability

The datasets in the article could be downloaded from the URL: https://github.com/zhooufeng/Data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research was supported by the National Science Foundation of China under Grant Nos. 61572225 and 61472049, the Foundation of Jilin Provincial Education Department under Grant No. JJKH20190724KJ, the Jilin Province Science & Technology Department Foundation under Grant Nos. 20190302071GX and 20200201164JC, and the Development and Reform Commission Foundation of Jilin province under Grant No. 2019C053-11.

References

[1] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based Arabic text categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020.

[2] X. Yu, S. Mayhew, M. Sammons, and D. Roth, "On the strength of character language models for multilingual named entity recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 2018.

[3] Y. Wang, Y. Sun, Z. Ma, L. Gao, and Y. Xu, "Named entity recognition in Chinese medical literature using pretraining models," *Scientific Programming*, vol. 2020, Article ID 8812754, 9 pages, 2020.

[4] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August 2016.

[5] C. Alt, A. Gabryszak, and L. Hennig, "Probing linguistic features of sentence-level representations in relation extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Washington, DC, USA, July 2020.

[6] Z.-X. Liu, D. G. Zang, G.-Z. Luo, M. Lian, and B. Liu, "A new method of emotional analysis based on CNN–BiLSTM hybrid neural network," *Cluster Computing*, vol. 23, pp. 2901–2913, 2020.

[7] B. Peng, J. Wang, and X. Zhang, "Adversarial learning of sentiment word representations for sentiment analysis," *Information Sciences*, vol. 541, pp. 426–441, 2020.

[8] M. Seok, H.-J. Song, C.-Y. Park, J.-D. Kim, and Y.-S. Kim, "Named entity recognition using word embedding as a feature," *International Journal of Software Engineering and Its Applications*, vol. 10, no. 2, pp. 93–104, 2016.

[9] J. Yu, X. Jian, H. Xin, and Y. Song, "Joint embeddings of Chinese words, characters, and fine-grained subcharacter components," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017.

[10] M. Yin, C. Mou, K. Xiong, and J. Ren, "Chinese clinical named entity recognition with radical-level feature and self-attention mechanism," *Journal of Biomedical Informatics*, vol. 98, Article ID 103289, 2019.

[11] N. Zhang, G. Xu, Z. Zhang, and F. Li, "MIFM: multi-granularity information fusion model for Chinese named entity recognition," *IEEE Access*, vol. 7, pp. 181648–181655, 2019.

[12] N. V. Sobhana, P. Mitra, and S. K. Ghosh, "Conditional random field based named entity recognition in geological text," *International Journal of Computer Applications*, vol. 1, no. 3, pp. 143–147, 2010.

[13] G. D. Zhou and S. Jian, "Named entity recognition using an HMM-based chunk tagger," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, July 2002.

[14] M. Zhang, Y. Zhang, and D. Tin Vo, "Neural networks for open domain targeted sentiment," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.

[15] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[16] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He, "Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition," *Journal of Biomedical Informatics*, vol. 92, Article ID 103133, 2019.

[17] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, http://arxiv.org/abs/1508.01991.

[18] M. Cho, J. Ha, C. Park, and S. Park, "Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition," *Journal of Biomedical Informatics*, vol. 103, Article ID 103381, 2020.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, June 2016.

[20] C. Wang, W. Chen, and B. Xu, "Named entity recognition with gated convolutional neural networks," in *Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 110–121, Nanjing, China, October 2017.

[21] S. Cao, W. Lu, J. Zhou, and X. Li, "cw2vec: learning Chinese word embeddings with stroke n-gram information," in *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 5053–5061, New Orleans, LA, USA, February 2018.

[22] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 2018.

[23] C. Xu, F. Wang, J. Han, and C. Li, "Exploiting multiple embeddings for Chinese named entity recognition," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Beijing, China, November 2019.

[24] Y. Wu, X. Wei, Y. Qin, and Y. Chen, "A radical-based method for Chinese named entity recognition," in *Proceedings of the 2nd International Conference on Big Data Technologies*, Jinan, China, August 2019.

[25] N. Ye, X. Qin, L. Dong, X. Zhang, and K. Sun, "Chinese named entity recognition based on character-word vector fusion," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8866540, 7 pages, 2020.

[26] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, "An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition," *IEEE Access*, vol. 7, pp. 113942–113949, 2019.

[27] J. Chung, K. Cho, and Y. Benjio, "A character-level decoder without explicit segmentation for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August 2016.

[28] M.-T. Luong and C. D. Manning, "Achieving open vocabulary neural machine translation with hybrid word-character models," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August 2016.

[29] Y. Xiao and K. Cho, "Efficient character-level document classification by combining convolution and recurrent layers," 2016, http://arxiv.org/abs/1602.00367.

[30] B. Liu, Y. Zhou, and W. Sun, "Character-level hybrid convolutional and recurrent neural network for fast text categorization," in *Proceedings of the International Conference on Extreme Learning Machine*, Singapore, November 2018.

[31] A. Chen, F. Peng, R. Shan, and G. Guo-Zheng Sun, "Chinese named entity recognition with conditional probabilistic models," in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia, July 2006.

[32] J. Zhou, H. Liang, D. Xinyu, and C. Jiajun, "Chinese named entity recognition with a multi-phase model," in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia, July 2006.

[33] J. Zhou, W. Qu, and F. Zhang, "Chinese named entity recognition via joint identification and categorization," *Chinese Journal of Electronics*, vol. 22, no. 2, pp. 225–230, 2013.

[34] C. Dong, J. Zhang, and C. Zong, "Character-based LSTM-CRF with radical-level features for Chinese named entity recognition," in *Proceedings of the International Conference on Computer Processing of Oriental Languages National CCF Conference on Natural Language Processing and Chinese Computing*, pp. 239–250, Kunming, China, December 2016.

[35] Y. Zhang and Z. Yang, "Chinese ner using lattice lstm," in *Proceedings of the The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.

[36] H. Zhao, M. Xu, and J. Cao, "Pre-trained language model transfer on Chinese named entity recognition," in *Proceedings of the 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, October 2019.

[37] S. Johnson, S. Shen, and Y. Liu, "CWPC_BiAtt: character-word-position combined BiLSTM-attention for Chinese named entity recognition," *Information*, vol. 11, no. 1, p. 45, 2020.

[38] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. 12, pp. 2493–2537, 2011.

[39] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[40] Y. Shen, H. Yun, Z. C. Lipton, and Y. Kronrod, "Deep active learning for named entity recognition," 2017, http://arxiv.org/abs/1707.05928.

[41] Y. ZhangS. Clark et al., "A fast decoder for joint word segmentation and POS-tagging using a single discriminative model," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT Stata Center, MA, USA, October 2010.

Hindawi
Complexity
Volume 2021, Article ID 6658516, 9 pages
https://doi.org/10.1155/2021/6658516

Research Article

An Evaluation Study on Investment Efficiency: A Predictive Machine Learning Approach

Weiwei Hao, Hongyan Gao, and Zongqing Liu

School of Economics and Management, Beijing Jiaotong University, Beijing, China

Correspondence should be addressed to Weiwei Hao; wwhao@bjtu.edu.cn

Received 26 December 2020; Revised 7 January 2021; Accepted 18 January 2021; Published 8 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

This paper proposes a nonlinear autoregressive neural network (NARNET) method for the investment performance evaluation of state-owned enterprises (SOE). It is different from the traditional method based on machine learning, such as linear regression, structural equation, clustering, and principal component analysis; this paper uses a regression prediction method to analyze investment efficiency. In this paper, we firstly analyze the relationship between diversified ownership reform, corporate debt leverage, and the investment efficiency of state-owned enterprises (SOE). Secondly, a set of investment efficiency evaluation index system for SOE was constructed, and a nonlinear autoregressive neural network approach was used for verification. The data of A-share state-owned listed companies in Shanghai and Shenzhen stock exchanges from 2009 to 2018 are taken as a sample. The experimental results show that the output value from the NARNET is highly fitted to the actual data. Based on the neural network model regression analysis, this paper conducts a descriptive statistical analysis of the main variables and control variables of the evaluation indicators. It verifies the direct impact of diversified ownership reform on the investment efficiency of SOE and the indirect impact on the investment efficiency of SOE through corporate debt leverage.

1. Introduction

Under the complete market hypothesis, the company's investment decisions are only related to investment opportunities, not to financing methods. However, due to information asymmetry, agency conflict, financing constraints, and other factors in the real market, the company's debt ratio will directly affect the company's investment decisions, resulting in that the company's actual investment scale deviates from the optimal investment scale, resulting in the inefficient investment situation of underinvestment or overinvestment.

Early studies began with the debt leverage and corporate investment efficiency from the agency conflict theory and contingency governance theory. From the agency conflict theory, the agency conflict between shareholders and creditors caused by liabilities often leads shareholders or managers to make investment decisions that damage the interests of creditors and then lead to investment distortion. Jensen and Meckling [1] and Myers [2] put forward the problem of low investment efficiency from the perspective of the conflict of interests between creditors and shareholders. In other words, in the financing structure with high financial leverage, managers, as one of the shareholders, have strong motivation to engage in high-risk investment projects. Starting from the theory of contingent governance, debt can reduce the agency costs caused by conflicts between shareholders and managers through the supervision and management by creditors, thereby effectively restraining managers from making investment decisions that damage shareholders' rights. Tirole [3] believed that debt constraints can urge companies to use free cash flow for investment projects with positive net present values, instead of being consumed by management at will or investing in projects with negative net present values. Managers who face liquidity shortages caused by debt constraints will be more cautious and careful in project investment, which will help to improve the company's investment efficiency.

The current analysis is mostly based on the premise of hard budget constraints. Sun and Li [4] examined the investment of research funds at universities and six variables were selected as evaluation indicators from the perspective

of fixed assets, teaching configuration, research instruments, and the number of books in libraries. Through empirical analysis, Li and others [5] believe that most of the nonfinancial listed companies in China have excessive investment behaviors, showing inefficiency of investment. The greater the absolute amount of annual increase in debt, the more serious the inefficiency of corporate investment. The smaller the ratio of debt to debt is, the more likely it is to cause underinvestment in the enterprise.

It had been proved that the latest developments in data science and machine learning have the potential to improve investment decisions. Many scholars have used machine learning methods to research and evaluate smart investment strategies. Li Liang and Wang Jia [6] analyzed the investment efficiency of China Merchants Bank's smart investment based on K-means cluster analysis and data mining technology and proved that the method is suitable for efficiency analysis of other smart investment products. The linear regression, decision tree, random forest, neural network, and XGBoost models were applied to analyze the efficiency of more than 10,000 companies in Li et al.'s work [7]. They proposed an interpretable machine learning algorithm for intelligent decision-making problems to predict the innovation efficiency of the enterprise [7]. Bhagchandani [8] used machine learning to reveal the gap between financial ecosystems and designed a basic algorithm to analyze the graphical structure of the financial market, using logistic regression to determine the validity of predictions.

Different scholars have done a lot of research on the investment efficiency of state-owned enterprises using machine learning methods from different perspectives, including exploring the impact on corporate investment efficiency from the debt and analyzing the impact on corporate investment efficiency from the perspective of diversified ownership reform. These have laid a good foundation for the study of this paper. However, the existing research on the relationship between diversified ownership reform, debt leverage, and investment efficiency of state-owned enterprises mainly focuses on the relationship between the two, and a few works have connected the above three.

Starting from the theory of capital structure, this paper analyzes the relationship among diversified ownership reform, corporate debt leverage, and investment efficiency of state-owned enterprises and establishes a comprehensive index evaluation system for investment efficiency. We used the diversified ownership reform and corporate debt leverage as impact factors, used the entropy method for data processing and analysis, used a neural network for training samples, and established an investment efficiency evaluation model based on NARNET.

We use NARNET to conduct regression analysis and evaluation on the investment efficiency of state-owned enterprises and draw conclusions about the influence of diversified ownership reform and corporate debt leverage on the investment efficiency of state-owned enterprises.

The second part of this paper gives an overview of the related research of machine learning applied in the field of efficiency evaluation. In the third part, we establish the investment efficiency evaluation index system, process the sample data set of state-owned listed companies, and establish the NARNET model. In the fourth part, the neural network is trained to achieve the desired effect and the regression analysis is performed on sample data. Finally, the conclusion of this paper is given in the fifth part.

2. Related Work

In recent years, machine learning has achieved many successes in the fields of image and face recognition, speech recognition, natural language processing, recommendation systems, and so forth. In terms of efficiency evaluation, the use of machine learning has also become a hot topic in related research.

Apipie and Georgescu [9] applied multiobjective optimization metaheuristics to solve a portfolio optimization problem. By analyzing the characteristics of listed real estate companies and the ideological basis of the support vector machine theory, they designed the evaluation index selection principle, formed the evaluation index system composed of ten indexes, and established the support vector machine regression model. Liu [10] proposed a performance auxiliary analysis system based on text analysis, using machine learning based on SVM, Naive Bayes, K-nearest neighbors, and other text classification algorithms to classify the roles of government officials, and established a quantitative role evaluation system. It provides a more objective and scientific reference for the establishment of the performance appraisal system on the e-government system based on the natural interaction mode.

Li proposed using SVM to solve the advantages of a small-sample, nonlinear, and high-dimensional pattern recognition to evaluate and analyze stock quantitative investment [11]. In Wu's work, the Stackelberg game model of green investment decision-making among enterprises is established by considering the case of the supplier's green investment alone and the case of the manufacturer and the supplier's joint green investment. The influence of green uncertainty on enterprise's decision-making is analyzed [12]. The previous studies' results proved that the machine learning method is in the performance evaluation system and the model has good applicability. Lin et al. put forward a new type of multiobjective optimization method for performance evaluation problems using the K-means algorithm to classify the evaluation objects and then construct the satisfaction function [13]. Based on the above work, they established a multiobjective optimization model to evaluate the satisfaction of the object. The results prove that the scalar model can obtain a weakly effective solution. Song [14] constructed a fuzzy chance-constrained least-squares twin support vector machine (FCC-LSTSVM) in business performance prediction. Ahmed [15] determined the standard rules of the best KPI for e-commerce websites based on Google Analytics and decision tree algorithm. Heilbrunn [16] used the machine learning method of the alternating diffusion process to study the impact of strategic planning on the performance of service-oriented small- and medium-sized enterprises (SMEs). This method applies two business data sets, which are strategic planning data and performance data.

Li and others constituted the evaluation index system of the fundamental innovation performance of the enterprise from the six dimensions of resources, technology, products, management, commercial value, and social value [17]. They established a performance evaluation model and trained the model applying BP neural network and finished the simulation verification of the fundamental innovation performance evaluation of sample enterprises. Wang used RS-Multi-Boosting as a hybrid integrated machine learning (HEML) method to help companies make correct business decisions and assessments between NPD (new product development) incremental strategies and aggressive NPD strategies to improve NPD's overall performance [18].

Different from the methods mentioned above, this paper does not directly apply a nonlinear recurrent neural network with external inputs method to efficiency evaluation. Because of the efficiency of state-owned enterprises as a complex system, direct data-based learning is not enough to have a better explanation. In this paper, we used the advantages of the entropy method in data processing to determine the weight of the initial weight of each indicator. The NARNET is used to evaluate the regression method of state-owned enterprise investment efficiency. Experiments show that the NARNET model is more consistent with data simulation. We verified the relationship between diversified ownership reform, corporate debt leverage, and other factors and the investment efficiency of state-owned enterprises. The results verify that this method is effective and applicable.

3. Method

3.1. Outline. The relationship among mixed ownership reform, corporate debt, and corporate efficiency is based on the following analysis.

Diversified ownership reforms have an impact on SOE debt mainly through the following ways: (1) Through diversified ownership reforms, non-state-owned capital is added, and the capital scale of enterprises is expanded. It can reduce the asset liability ratio of state-owned enterprises when the total debt scale of enterprises remains unchanged. At the same time, enterprises can also use this part of the expanded funds to give priority to debt repayment, so as to reduce the debt ratio of enterprises and achieve the purpose of reducing debt. (2) Diversified ownership reforms can indirectly reduce the debt of SOE by improving corporate governance. By improving the corporate governance system and strengthening the market-oriented operation, shareholder supervision will be more adequate, which will change the situation of excessive debt of enterprises, and promote enterprises to return to a healthy level of asset liability ratio.

The investment efficiency of state-owned enterprises is mainly affected by the following ways. The introduction of non-state-owned capital by SOE through diversified ownership reforms can play a supervisory role in the management, so that the management can make more scientific and reasonable investment decisions, reduce ineffective and irrational investment behavior, and improve the investment efficiency of state-owned enterprises.

Through the above analysis, diversified ownership reforms are conducive to reducing the debt of SOE, and reducing the debt of state-owned enterprises is conducive to improving the investment efficiency of enterprises. Furthermore, it can be seen that diversification can not only directly affect the efficiency of investment but also indirectly affect the efficiency of investment.

The investment efficiency of enterprises is a complex system. However, the existing research on investment efficiency is from an explanatory point of view, which is lack of comprehensiveness. In this paper, we propose a predictive method to test our hypothesis through a regression neural network.

The process of building a neural network model is given in Figure 1. As shown in Figure 1, we first choose the evaluation index of investment efficiency, which may directly or indirectly affect the investment efficiency. We complete this work through empirical method. Next, we use entropy method to determine the weight of each index. The above work can only explain the relationship between each index and investment efficiency. Finally, we train a regression neural network model. We use predictive methods to verify the comprehensive relationship between these indicators and investment efficiency.

3.2. Construction of the Model Indicator System. To build a comprehensive evaluation model of the investment efficiency of state-owned enterprises, we should first determine the model index according to the theoretical basis of enterprise debt and investment.

The explained variable in this study is investment efficiency. This paper draws on Richardson's [19] expectation model to estimate investment efficiency and uses the absolute value of the residual estimated by the regression model to measure inefficient investment. With lower efficiency, the residual value greater than zero indicates overinvestment, and the residual value less than zero indicates underinvestment. Therefore, the higher the degree of inefficient investment, the lower the investment efficiency.

3.2.1. Total Investment. $\mathrm{Inv}_{i,t}$ represents the total investment of company i in year t:

$$\mathrm{Inv}_{i,t} = \frac{\text{capital expenditure of } i}{\text{total assets at the end of } t} \times 100\%. \quad (1)$$

3.2.2. Investment Opportunities. $\mathrm{Grow}_{i,t-1}$ represents investment opportunities, expressed by the growth rate of company i operating income in year $t-1$.

3.2.3. Liabilities. $\mathrm{Lev}_{i,t-1}$ represents company i debt situation in year $t-1$, expressed in terms of asset-liability ratio.

3.2.4. Current Cash Situation. $\mathrm{Cash}_{i,t-1}$ indicates the current cash situation of a company i in year $t-1$:

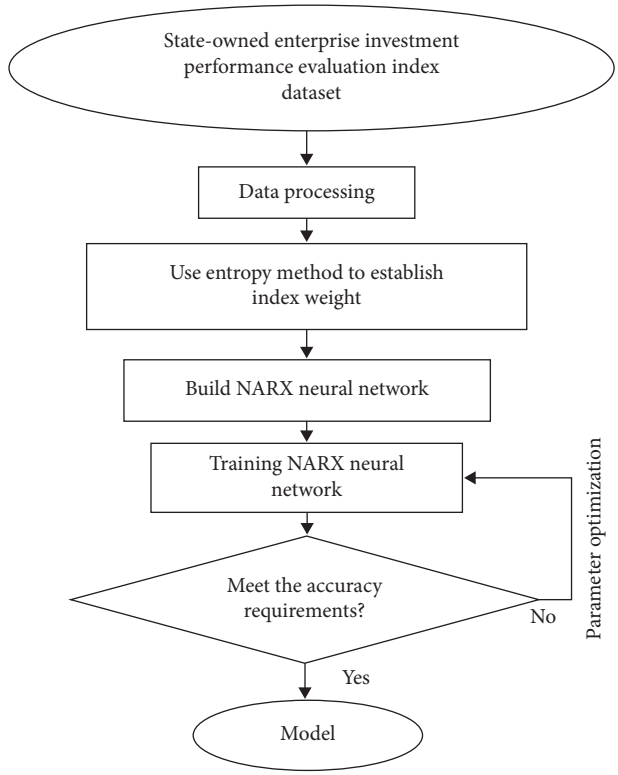


Figure 1: Flow chart of neural network establishment.

$$\text{Cash}_{i,t-1} = \frac{\text{currency assets } i \text{ at the end of } t-1}{\text{total assets at the end of } t-1} \times 100\%. \tag{2}$$

3.2.5. Listing Years. $\text{Age}_{i,t-1}$ indicates the length of company i listing period at the end of period $t-1$.

3.2.6. Company Size. $\text{Size}_{i,t-1}$ represents the size of the company; the calculation method is that the company's total assets at the end of the year take the natural logarithm:

$$\text{Size}_{i,t-1} = \log(\text{company total assets at the end of the year}). \tag{3}$$

3.2.7. Profitability. $Ar_{i,t-1}$ indicates the profit level of company i in year $t-1$, expressed by basic earnings per share.

3.2.8. Dummy Variables. We use Ind and Year as dummy variables of industry and year to remove the influence caused by different years and industries.

To comprehensively measure the degree of diversified ownership reform of state-owned enterprises, four variables in the process of diversified ownership reform of state-owned enterprises (Var), depth (Dep), degree of integration (Mix), and degree of checks and balances (Res) are selected as the diversified ownership reform explanatory variables.

We selected the interest-bearing debt ratio indicator (Debr) to express the debt leverage of the company [20], only considering two ways of loans and bonds and using total assets for standardized treatment.

As explained above, this paper selects the following control variables: company size (Size), which is the natural logarithm of the total assets at the end of the period; cash holdings (Cash), which are the total monetary capital holdings divided by the total assets at the end of the period; and asset-liability ratio (Lev), which is the total liabilities at the end of the period divided by the total assets at the end of the period. Besides, this paper also eliminates the effects of annual dummy variables (Year) and industry dummy variables (Industry).

3.3. Data Sample Processing. The research in this paper selects the relevant data of A-share listed companies in Shanghai and Shenzhen stock exchanges from 2009 to 2018 as the initial sample, which was processed according to the following criteria:

(1) Selecting A-share state-owned listed companies in Shanghai and Shenzhen stock exchanges according to the nature of the actual controller;

(2) Excluding enterprises that belong to the financial industry;

(3) Excluding ST and *ST companies. The operating conditions of these companies are deteriorating, and some of their financial indicators deviate from the normal listing requirements. Therefore, the risk of

taking these companies as samples will affect the results more often;

(4) Excluding sample companies that have been listed for less than 5 years to weaken the impact of abnormal changes in investment efficiency caused by listing;

(5) Eliminating companies with missing relevant indicator data. To control the influence of extreme values on the results of the study, a 1% Winsorized extreme value processing was performed at both ends of the data.

At the same time, this paper selects 10% as the cut-off point and excludes the case where the non-state-owned constituents hold no more than 10% of the shares in state-owned enterprises [21]. 393 samples and 3930 observations were finally obtained.

Due to the large actual value of asset-related index data, the network training convergence will be slow and the training time will be too long. That will lead to the bad efficiency of the model. Therefore, it is required to normalize the larger data. This paper uses the max-min method for processing:

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{4}$$

To provide a basis for the comprehensive evaluation of multiple indicators, this paper uses the entropy method to determine the initial weight of each indicator. The process is as follows:

(1) Establishing an indicator data matrix:

$$A = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}_{n*m}. \tag{5}$$

For a sample of n enterprises and m indicators, X_{ij} is the value of the j-th indicator of the enterprise i($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$).

(2) Data standardization:

The same max-min method as above is used to perform data translation processing on the indicator.

(3) It requires the proportion of the j-th index value of the i-th company to calculate the index entropy.

$$y_{ij} = \frac{x'_{ij}}{\sum_{i=1}^{n} x'_{ij}}. \tag{6}$$

(4) To calculate the information entropy of the j-th index, the formula is

$$e_j = -k \sum_{i=1}^{m} y_{ij} \ln y_{ij}, \tag{7}$$

k is a constant. $k = (1/\ln m)$.

(5) Calculating the difference coefficient of the j-th index:

$$d_j = 1 - e_j. \tag{8}$$

(6) The weight of the jth index is

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, \quad j = 1, 2 \ldots m. \tag{9}$$

Finally, $W = \{w_1, w_2, \ldots, w_m\}$ is the objective weight value of each indicator.

3.4. NARNET. In this paper, NARNET is a nonlinear autoregressive neural network with external inputs. A typical NARNET is mainly composed of the input layer, hidden layer, output layer, and input delay function. The basic structure of the model is shown in Figure 2.

The difference between NARNET and an ordinary neural network is that the former adds a delay function before the hidden layer. The output layer of the neural network receives feedback from the hidden layer, which can be described as

$$y(t+1) = f[y(t), y(t-1), \ldots, y(t-ny+1), u(t), u(t-1), \ldots, u(t-nu+1, W)] = f[y(t), u(t), W]. \tag{10}$$

In the NARNET, the output signal is delayed and then is input into the neural network; through the hidden layer and the output layer, the final output result is obtained. Taking i as the amount of input data and j as the number of hidden layer neurons, x_i represents the i-th input signal of the network, and w_{ij} represents the connection weight between the i-th output delay signal and the j-th neuron. a_j represents the threshold value of the j-th hidden layer neuron, and the hidden layer activation function f is synthesized to obtain the calculation results of each neuron:

$$d_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + a_j\right). \tag{11}$$

Then the model takes w_j as the connection weight between the j-th neuron in the hidden layer and the output layer neuron and b as the output layer neuron threshold. The output result is calculated as follows:

$$o_j = f\left(\sum_{j=1}^{n} w_{ij} d_i + b\right). \tag{12}$$

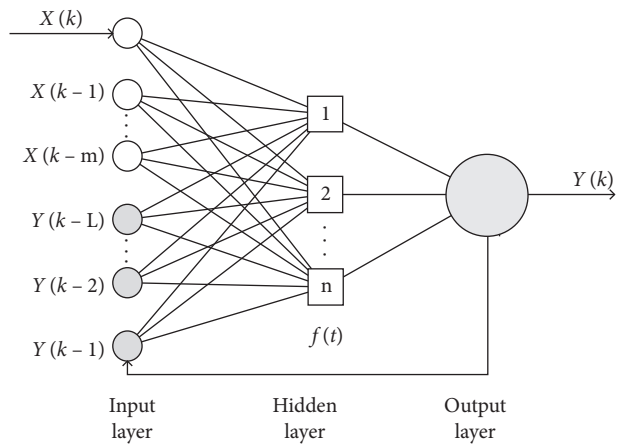


Figure 2: The structure of the NARNET.

4. Results and Discussion

According to the above training data set, the entropy method is used to calculate the weight of each evaluation index, as shown in Figure 3. Preliminary analysis shows that the contributions of investment indicators, diversified ownership reform, and corporate debt leverage to the evaluation model are about 3/5, 1/5, and 1/5, respectively.

According to the evaluation index system established in the previous section, we collect x1 (total investment volume), x2 (investment opportunities), x3 (debt situation), x4 (liquid cash situation), x5 (length of listing period), x6 (company size), x7 (profit level), x8 (dummy variable (industry)), x9 (dummy variable (year)), x10 (equity diversity), x11 (equity depth), x12 (equity mixing degree), x13 (equity balance), and x14 (interest-bearing debt ratio) to constitute the enterprise investment efficiency evaluation data table in Table 1.

This paper takes the first indicator in the model evaluation indicator system as the output and the remaining indicators as the input. The BP algorithm was taken as the network training method and the number of hidden layer neurons is set to 14, and the delay is set to 2. It can be seen from Figure 4 that the network has reached the setting accuracy and has a good fitting to the sample data. So the training of the model is ideal. Figure 5 shows the error autocorrelation function of this network. It can be seen from Figure 6 that the error autocorrelation function is within the confidence interval except for individual special data. So the network is well trained.

It can be seen from Figure 4 that the output value from the NARNET is highly fitted to the actual data. Based on the neural network model's regression analysis, this paper conducts a descriptive statistical analysis of the main variables and control variables of the evaluation indicators. It verifies the direct impact of diversified ownership reform on the investment efficiency of state-owned enterprises and the indirect impact on the investment efficiency of state-owned enterprises through corporate debt leverage. The model independently estimates the impact of state-owned enterprise debt leverage on enterprise investment efficiency, and the results are given in Table 2.

The results show that corporate debt leverage and corporate investment efficiency are in an inverted U-shaped relationship. Initially, as corporate debt leverage increases, corporate investment efficiency will increase. After reaching a certain peak, as corporate debt leverage continues to increase, the level of investment efficiency began to decline. This shows that corporate debt leverage has a certain limit; it is not that the higher the better.

State-owned enterprises not only undertake economic goals such as promoting economic growth but also bear the responsibility of promoting employment and maintaining social stability; therefore, inefficient investment was often made in these companies. Chen and Dong [22] believe that the government's distorted market mechanism has hindered the development of the state-owned economy. Du et al. [23] have shown that political connections will reduce the efficiency of corporate investment. State-owned enterprises introduced nonstate capital through the implementation of mixed-ownership reform to optimize the ownership structure. This is helpful for enterprises to make correct business decisions, reduce inefficient investment behavior, and improve the investment efficiency of state-owned enterprises. Besides, the diversified ownership reform has expanded the capital scale of enterprises. Therefore, the debt ratio of the enterprise can be reduced, and the purpose of deleveraging can be achieved. The addition of nonpublic components dilutes the state-owned components, which increases the debt cost of diversified-ownership enterprises to a certain extent, thereby reducing the debt leverage ratio of enterprises.

According to the trade-off theory, appropriate debt will bring tax benefits to the enterprise. When the company's asset-liability ratio is low, the company can moderate debt to ease the financing constraints of the company and meet the needs of the company's investment expansion, thereby curbing the underinvestment of the company and improving the efficiency of the company's investment. When the cost of debt is greater than the mitigation effect of debt financing on corporate financing constraints, this will increase the cost of

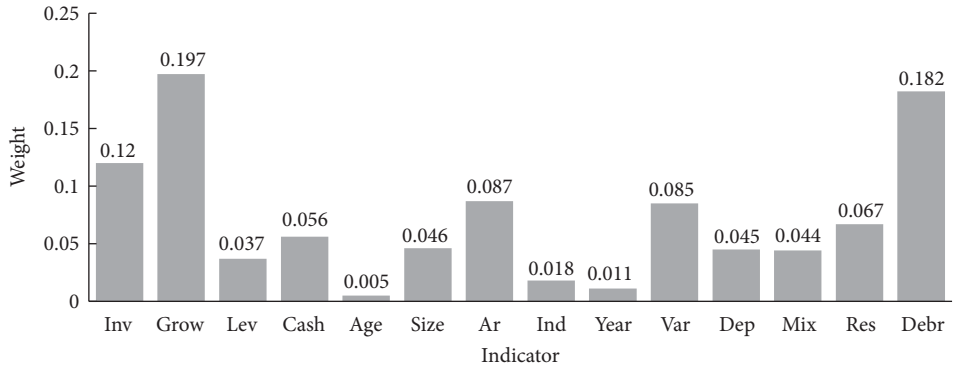


Figure 3: Indicator weight diagram.

Table 1: Investment efficiency evaluation data sheet.

	Sample				
Indicator	1	2	3	...	4324
x1	0.1805	0.2493	0.0695		12.0348
x2	0.1538	−0.0022	0.6788		0.0029
x3	0.6744	0.7166	0.7435		0.4568
x4	0.1675	0.0263	0.1649		0.1398
x5	29				13
x6	0.255				0.237
x7	0.4219				0.6609
x8	1				12
x9	2009				2009
x10	0.7166				0.4896
x11	0.7435		...		0.517
x12	0.0695				0.7015
x13	0.1649				0.3108
x14	0.1398				0.6071

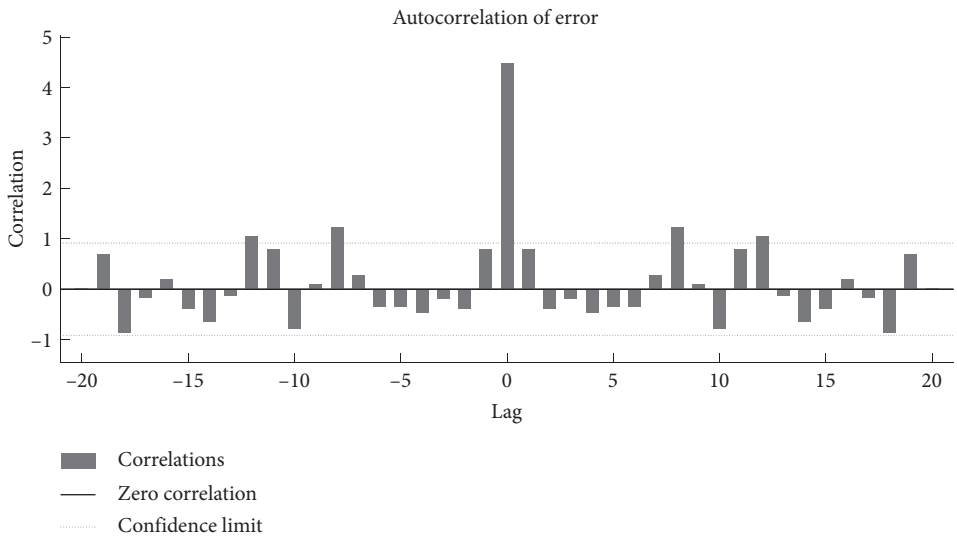


Figure 4: Autocorrelation error.

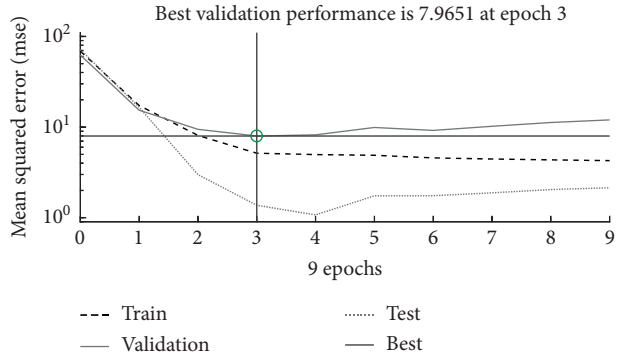


Figure 5: Neural network training iteration process.

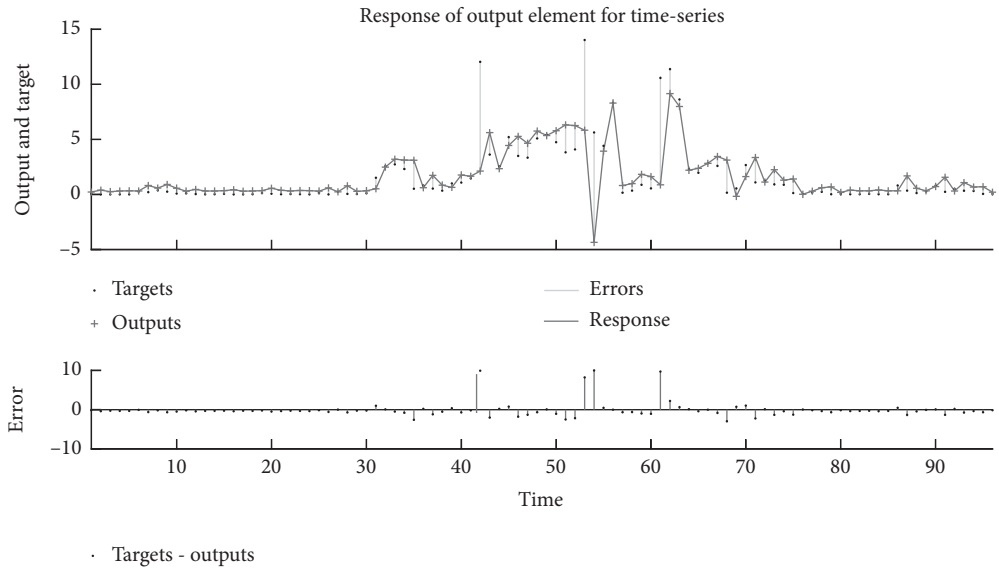


Figure 6: Diagram of neural network fitting results

Table 2: Regression debt leverage on enterprise investment efficiency model.

Variable	Debt	Debt^2	Size	Cash	Grow	Constant	R^2	Prob > F
Noinvest	1.446**	—	0.104	−1.657	−0.073	−2.611	0.0042	0.0053
	3.856**	4.428*	0.281	−2.287**	−0.135	−6.535	0.0063	0.0000

supervision by creditors, leading to excessive corporate investment.

5. Conclusion

Through the above theoretical analysis and model verification of the relationship between state-owned enterprise debt leverage, diversified ownership reform, and investment efficiency, it is shown that, firstly, the diversified ownership reform can effectively improve the diversity of equity, the degree of equity integration, the degree of equity integration, and the degree of equity balance. It can also reduce the debt leverage of state-owned enterprises. Secondly, the debt leverage of state-owned enterprises and enterprise investment efficiency show a significant inverted U-shaped relationship. Furthermore, the diversified ownership reform can not only directly affect the investment efficiency of state-owned enterprises but also indirectly affect the investment efficiency of state-owned enterprises by acting on corporate debt leverage; at the same time, there is a time lag. Therefore, in the process of diversified reform of state-owned enterprises, we must pay attention to the proportion of equity to be introduced while actively introducing multiple ownership equity to make sure it can play a substantial role in the various production and operation decisions of the enterprise.

In this paper, the NARNET is used to evaluate the regression method of state-owned enterprise investment efficiency. Experiments show that the network model is more consistent with data simulation. We verified the relationship between diversified ownership reform, corporate debt leverage, and other factors and the investment efficiency of state-owned enterprises. The results verify that this method is effective and applicable.

The investment efficiency evaluation of SOE is a complex problem. The shortcomings of this paper are mainly in the stage of data processing. We use the entropy method to weigh the data indicators. We do the above processing because we think that the impact of the above indicators on the investment efficiency is different. However, we did not test this hypothesis. In addition, this paper represents a regression analysis method, and we only use our method for comparison with the actual one. More experiments will be the next priority of this paper. In addition, we will consider more factors and technologies to be used to evaluate investment efficiency, such as investment strategy [24, 25], big data [26], and hybrid machine learning [27, 28] methods in the future.

Data Availability

All the data used in this study can be obtained upon request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. C. Jensen and W. H. Meckling, "Theory of the firm: managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, vol. 3, no. 4, pp. 305–360, 1976.

[2] S. T. C. Myers, "Determinants of corporate borrowing," *The Journal of Finance*, vol. 5, pp. 147–175, 1977.

[3] J. Tirole, *The Theory of Corporate Finance*, Princeton University Press, Princeton, NJ, USA, 2006.

[4] J. Sun, Y. Li, X. Zhao, and N. Zhang, "An evaluation on investment of research funds with a neural network algorithm in "double first-class" universities," *Complexity*, vol. 2020, Article ID 7496126, 8 pages, 2020.

[5] L. Qiang, J. Ji, and H. Ju, "Inefficient investment and debt structure: empirical evidence from China," *Investment Research*, vol. 3, pp. 66–79, 2014.

[6] L. Li, J. Wang, and X. Li, "Efficiency analysis of machine learning intelligent investment based on K-means algorithm," *IEEE Access*, vol. 8, pp. 147463–147470, 2020.

[7] Y. Li, L. Yang, B. Yang, N. Wang, and T. Wu, "Application of interpretable machine learning models for the intelligent decision," *Neurocomputing*, vol. 333, pp. 273–283, 2019.

[8] A. Bhagchandani and D. Trivedi, "A machine learning algorithm to predict financial investment," in *Data Science and Intelligent Applications*, pp. 261–266, Springer, Singapore, 2020.

[9] F.-M. Apipie and V. Georgescu, "Assessing and comparing by specific metrics the performance of 15 multiobjective optimization metaheuristics when solving the portfolio optimization problem," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 3/2019, pp. 39–58, 2019.

[10] S. Liu, "Auxiliary performance evaluation system based on text analysis," Dissertation, Harbin Institute of Technology, Harbin, China, 2015.

[11] Li Jiang, "Research on stock quantitative timing strategy based on support vector machine," Dissertation, Shenzhen University, Shenzhen, China, 2019.

[12] S. Wu, X. Yao, and G. Wu, "Environmental investment decision of green supply chain considering the green uncertainty," *Complexity*, vol. 2020, Article ID 8871901, 13 pages, 2020.

[13] Y. Lin, W. Liu, and Y. Wang, "An integrated approach using cross-efficiency and shapley value in performance evaluation," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 4, pp. 209–224, 2019.

[14] Y.-G. Song, Q.-L. Cao, and C. Zhang, "Towards a new approach to predict business performance using machine learning," *Cognitive Systems Research*, vol. 52, pp. 1004–1012, 2018.

[15] H. Ahmed, T. A. Jilani, W. Haider, M. A. Abbasi, S. Nand, and S. Kamran, "Establishing standard rules for choosing best KPIs for an e-commerce business based on google analytics and machine learning technique," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, pp. 12–24, 2017.

[16] S. Heilbrunn, N. Rabin, and S. Rozenes, "Detecting mutual configurations of applied planning strategies and performances in small and medium sized businesses with kernel based machine learning methods," *Applied Soft Computing*, vol. 61, pp. 1211–1225, 2017.

[17] H. Li, Q. Zhang, and Z. Zheng, "Research on enterprise radical innovation based on machine learning in big data background," *The Journal of Supercomputing*, vol. 76, no. 5, pp. 3283–3297, 2020.

[18] X. Wang, D. Zeng, H. Dai, and Y. Zhu, "Making the right business decision: forecasting the binary NPD strategy in Chinese automotive industry with machine learning methods," *Technological Forecasting and Social Change*, vol. 155, Article ID 120032, 2020.

[19] S. Richardson, "Over-investment of free cash flow," *Review of Accounting Studies*, vol. 11, pp. 159–189, 2006.

[20] H. Ma and Y. Wang, "Can deleveraging improve the investment efficiency of enterprises?—an empirical analysis based on the empirical data of Chinese listed companies," *Securities Market Herald*, vol. 5, pp. 13–20, 2017.

[21] H. Yang and L. Gong, "State-owned and private-owned diversified equity participation and company performance improvement," *Economic Research*, vol. 3, pp. 122–135, 2017.

[22] D. Cheng and Y. Dong, "Study on the productivity measurement and change trend of China's diversified ownership economy," *Economics and Management Research*, vol. 6, pp. 33–43, 2014.

[23] X. Du, Q. Zeng, and Y. Du, "Political connection, over-investment and corporate value: empirical evidence based on state-owned listed companies," *Financial Research*, vol. 8, pp. 93–110, 2011.

[24] T. K. Lee, J. H. Cho, D. S. Kwon, and S. Y. Sohn, "Global stock market investment strategies based on financial network indicators using machine learning techniques," *Expert Systems with Applications*, vol. 117, pp. 228–242, 2019.

[25] A. Wilinski and B. Kovalerchuk, "Visual knowledge discovery and machine learning for investment strategy," *Cognitive Systems Research*, vol. 44, pp. 100–114, 2017.

[26] T. A. Borges and R. F. Neves, "Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods," *Applied Soft Computing*, vol. 90, Article ID 106187, 2020.

[27] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2/2019, pp. 77–92, 2019.

[28] I. Stoica, "Solving system problems with machine learning," *Studies in Informatics and Control*, vol. 28, no. 2, pp. 119–132, 2019.

Hindawi
Complexity
Volume 2021, Article ID 5574732, 12 pages
https://doi.org/10.1155/2021/5574732

Research Article

Optimizing Ontology Alignment through Linkage Learning on Entity Correspondences

Xingsi Xue,[1,2,3] **Chaofan Yang**,[1,2,3] **Chao Jiang**,[1,2,3] **Pei-Wei Tsai**,[4] **Guojun Mao**,[2] **and Hai Zhu**[5]

[1]*Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, Fujian 350118, China*
[2]*School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou, Fujian 350118, China*
[3]*Intelligent Information Processing Research Center, Fujian University of Technology, Fuzhou, Fujian 350118, China*
[4]*Department of Computer Science and Software Engineering, Swinburne University of Technology, John Street, Hawthorn, Victoria 3122, Australia*
[5]*School of Network Engineering, Zhoukou Normal University, Zhoukou, Henan 466001, China*

Correspondence should be addressed to Xingsi Xue; jack8375@gmail.com

Received 10 January 2021; Revised 25 January 2021; Accepted 27 January 2021; Published 5 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

Data heterogeneity is the obstacle for the resource sharing on Semantic Web (SW), and ontology is regarded as a solution to this problem. However, since different ontologies are constructed and maintained independently, there also exists the heterogeneity problem between ontologies. Ontology matching is able to identify the semantic correspondences of entities in different ontologies, which is an effective method to address the ontology heterogeneity problem. Due to huge memory consumption and long runtime, the performance of the existing ontology matching techniques requires further improvement. In this work, an extended compact genetic algorithm-based ontology entity matching technique (ECGA-OEM) is proposed, which uses both the compact encoding mechanism and linkage learning approach to match the ontologies efficiently. Compact encoding mechanism does not need to store and maintain the whole population in the memory during the evolving process, and the utilization of linkage learning protects the chromosome's building blocks, which is able to reduce the algorithm's running time and ensure the alignment's quality. In the experiment, ECGA-OEM is compared with the participants of ontology alignment evaluation initiative (OAEI) and the state-of-the-art ontology matching techniques, and the experimental results show that ECGA-OEM is both effective and efficient.

1. Introduction

Semantic Web (SW) is proposed by Tims Berners-Lee in 1998, which makes the intelligent applications be able to understand a word's meaning in semantic level. Ontologies are the solution to the issue of data heterogeneity on SW since it is able to make consensus of a certain conception meaning of a field and provide abundant domain knowledge and semantic vocabulary for the interaction between application systems. However, due to SW's scattering essence, there might be different definitions on a concept in separate ontologies, which leads to the issue of ontology heterogeneity [1]. Ontology matching is regarded as an effective method to address it, and swarm intelligent algorithm- (SIA-) based ontology matching techniques have achieved good performance in past studies [2], such as genetic algorithm (GA) [3], particle swarm optimization algorithm (PSO) [4], firefly algorithm (FA) [2], and artificial bee colony algorithm (ABC) [5]. However, there are two drawbacks in the existing SIA-based approaches: (1) massive time and memory consumption is required, which heavily blocks the efficiency of the ontology matching process; (2) an expert of related field or a reference alignment is required in the process of ontology matching which is usually not available in real application conditions. To overcome these drawbacks, an extended

compact genetic algorithm-based ontology entity matching technique (ECGA-OEM) is proposed in this work, which uses both compact encoding mechanism [6, 7] and linkage learning approach to efficiently match the ontologies. In particular, our contributions are as follows:

(i) A new evaluating metrics on the ontology alignment is proposed, which is able to work without the reference alignment and the domain experts.

(ii) An optimal model on ontology entity matching problem is constructed.

(iii) An ECGA-OEM is proposed, which uses the linkage learning and compact encoding mechanism to efficiently address the ontology entity matching problem.

The rest of this paper is organized as follows: the related works are narrated in Section 2; the statement of ontology, ontology matching, and similarity measures are presented in Section 3; the ontology entity matching through ECGA proposed by this paper is revealed in Section 4; Section 5 presents the experiment results; and finally, Section 6 draws the conclusion and presents the future work.

2. Related Works

Ontology matching is a complex, time-consuming, and error-prone work, especially when the scale of ontologies is large. Recently, a number of the machine learning (ML) techniques [8–14] have been proposed to automatically determine the ontology alignment. To improve the matching efficiency, Araújo et al. [15] presented the matching system through parallel computing (PC) technique and Amin et al. [16] matching ontology based on cloud computing (CC). At the same time, SIA-based technique has achieved great performance in the ontology matching [1, 2, 17–20] domain [21–26].

Generally, ontology matching techniques are classified into two categories: ontology metamatching techniques and ontology entity matching techniques [27]. The former dedicates to address the problem that how to aggregate different similarity measures with appropriate weights, and the latter tries to directly determine the entity correspondence set between two ontologies. The first SIA-based ontology metamatching system is genetics for ontology alignment (GOAL), which aims at optimizing the aggregating weight set for different matchers [3, 28–30]. Memetic algorithms (MAs), which introduce local search (LS) strategy into evolutionary algorithms (EAs) to improve its local optimization capability, are proposed to solve ontology metamatching problem [31]. To overcome the drawback of overreliance reference alignment, Xue et al. presented a partial reference alignment (PRA), in which only a part of standard reference is used to assess the quality of alignment [32]. Furthermore, Xue and Wang proposed an innovative metric named unanimous improvement ratio (UIR) to assess the alignment's quality, in which the reference alignment is not required [33]. Besides, artificial bee colony (ABC) algorithm is also adopted to address ontology metamatching problem, which further improves the solution's quality [5].

During the matching process, ontology metamatching techniques need to maintain several similarity matrices, which leads to huge memory consumption. For this reason, ontology entity matching techniques, which aims at directly determining the optimal pair set, attracts the expert's interests. Genetic algorithm-based ontology matching (GAOM) firstly regards certain matching pairs set as the optimizing objective [34]. MA is also utilized to solve the ontology entity matching problem, whose performance outperforms GA [35]. Bock et al. [4] use PSO to solve the ontology entity matching. In detail, it evaluates the fitness of chromosomes through a certain aggregation strategy for multiple objective functions. Alves et al. [36] argue that instances consisting in the ontology can be used to improve alignment in the condition that knowledge is embedded in them. For this reason, Xue et al. [37] take also instance-level matching into consideration to further improve the quality of alignment.

3. Preliminaries

3.1. Ontology and Ontology Matching

Definition 1 (ontology). An ontology O is a 5-tuple [33]. $O = (C, P, I, \Lambda, \Gamma)$ where C is a set of classes that cannot be empty, P is a set of properties that cannot be empty, I is a set (it could be empty) of individuals that represent the instances of classes in the real world, Λ is a nonempty set of axioms that are used to check the consistency of ontologies or deduce new information, and Γ is a set of annotations that provide information metadata so that the researcher can understand. Particularly, C, P, and I make up the entities in ontologies.

Definition 2 (ontology matching). The ontology matching can be considered as a function $f(O_1, O_2, A', p, r)$, where O_1 and O_2 are two ontologies to be matched; A' is an existing initial alignment of O_1 and O_2; p is a set of parameters, e.g., threshold, in the process of ontology matching; and r is a set of external resources, e.g., background knowledge based and dictionaries, which assisted in ontology matching. The process of ontology matching is depicted in Figure 1, where A is the obtained ontology alignment.

An example of matching two ontologies is presented in Figure 2, where O_1 and O_2 are two ontologies to be matched in this figure. The strings in the rounded rectangle are the classes, e.g., "Reference," "Entry," and "Book." The black lines between two classes of the same ontology represent their relation "has a" or "is a" in turn; e.g., "Reference" has a "Book" and "Book" is a "Reference," which means "Reference" is the supclass of "Book," and "Book" is the subclass of "Reference." There are datatype properties that describe the features of a class; e.g., "Data," "Title," and "Human Creator" are the properties of "Reference." The instances of a class are in a rectangle; e.g., "Of Natures Obvious Laws & Processes in Vegetation" is an instance of "Article." The relation of the entities between the two ontologies are linked by the lines with double arrowhead, and there are symbols: " $\equiv$," "$\subseteq$" (or "$\supseteq$"), and "$\perp$," which, respectively, means equivalence, more

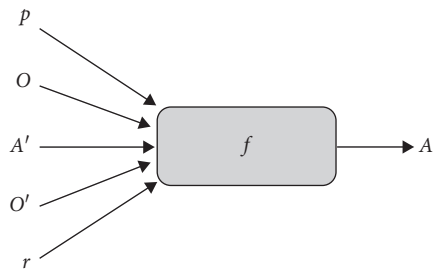


Figure 1: The process of ontology matching.

specific (or less specific), and disjointness relation; e.g., classes "Entry" and "Book" of O_2 are the equivalent and hyponym of class "Reference" of O_1, property "Event" of O_1, and property "Year" of O_2 are irrelevant.

3.2. Similarity Measures. It is necessary to measure to what extend two ontology entities are similar before finding the reliable entity correspondences. In the ontology domain, we usually use the similarity measure to calculate two entities' similarity values. Generally, there are three categories of similarity measures, i.e., syntactic, linguistic, and taxonomy-based measures [33].

3.2.1. Syntactic Measures. Two syntactic measures, i.e., SMOA (string metric for ontology alignment) [37] and Levenshtein [38], are employed in this paper. Given two strings s_1 and s_2, the SMOA and Levenshtein similarity are, respectively, defined as follows:

$$\mathrm{SMOA}(s_1, s_2) = \mathrm{Comm}(s_1, s_2) - \mathrm{Diff}(s_1, s_2) + \mathrm{WinklerImpr}(s_1, s_2), \tag{1}$$

where $\mathrm{Comm}(s_1, s_2)$ stands for the common length of s_1 and s_2 while $\mathrm{Diff}(s_1, s_2)$ for the different lengths and $\mathrm{WinklerImpr}(S_1, S_2)$ is the improvement to results yielded by the method that proposed by Winkler [39].

$$\mathrm{Levenshtein}(s_1, s_2) = \max\left(0, \frac{\min\left(|s_1|, |s_2|\right) - d(s_1, s_2)}{\min\left(|s_1|, |s_2|\right)}\right), \tag{2}$$

where $|s_1|$ and $|s_2|$ are the cardinality of the letters contained in s_1 and s_2, respectively, and $d(s_1, s_2)$ is the number of letters that need to be modified from s_1 to s_2. The final syntactic similarity is equal to the average of SMOA and Levenshtein.

3.2.2. Linguistic Measures. The linguistic similarity between two strings is worked out by considering the semantic relations (such as synonyms and hypernym) which usually requires using the thesaurus and dictionaries. In this work, WordNet [23, 40], an electronic vocabulary database that has collected every meaning of various words, is used. Given two words w_1 and w_2, Linguisitc Similarity (w_1, w_2) equals:

(i) 1, if words w_1 and w_2 are synonyms in Wordnet.

(ii) 0.5, if word w_1 is the hypernym of word w_2 or vice versa in Wordnet.

(iii) 0, otherwise.

3.2.3. Taxonomy-Based Measures. The core ideal of taxonomy-based measures is to make full use of the hierarchy relationship of ontology to determine two entities' similarity by considering their neighbor's similarity. In this work, a mutual reasoning between class and property (MRCP) is proposed as the taxonomy-based measure, which is shown in Figure 3.

In Figure 3, the circle is the class of the ontology, the triangle is the properties of the ontology, and the one-way arrow represents the hierarchical relationship; i.e., class c_{a1} is the supclass of class c_{a3}, the dividing line arrow between the class and the property indicates that the property belongs to this class, the bidirectional arrow indicates that there is a high similarity between the two entities, and the dashed two-way arrow indicates that the similarity between them is improved after the operation. Subgraph (a) depicts the classes' similarity gained from their neighbor, i.e., supclass and subclass. There is high similarity between classes c_{a1} and c_{b1}, so the similarity between their subclasses c_{a3} and c_{b2} is supposed to increase. Likewise, similarity of classes c_{a3} and c_{b2} would be increased because their subclasses c_{a6} and c_{b4} are highly similar. Subgraph (b) is the properties' similarity gained from their supproperty and subproperty. The similarity of properties p_{a3} and p_{b2} will be improved since their supproperties p_{a1} and p_{b1} and subproperties p_{a6} and p_{b4} are highly similar, respectively. Subgraph (c) is the properties' similarity gained from the classes which they belong to. c_{a1} and c_{b1} are the classes of two ontologies and p_{a1}, p_{a2}, p_{a3}, p_{b1}, p_{b2}, and p_{b3} are their properties, respectively. The similarities between properties of class c_{a1} and properties of class c_{b1} would be improved because of the high similarity of c_{a1} and c_{b1}; i.e., the similarity of p_{a3} and p_{b1} would be promoted and so are the remaining eight combinations. On the contrary, classes' similarity will be increased due to the same or highly similar properties they shared, as is depicted in subgraph (d). Since pairs p_{a1} and p_{b3}, p_{a2} and p_{b2}, and p_{a3} and p_{b1}, the similarity of c_{a1} and c_{b1} is increased too.

3.2.4. Aggregation Strategy. Three similarity matrixes are generated when the three measures have been applied. In this work, three matrices need to be aggregated into one matrix. The final similarity value $S_a(s_1, s_2)$ between two entities s_1 and s_2 is defined as follows:

$$S_a(s_1, s_2) = \begin{cases} 1, & S_s = 1 \text{ or } S_l = 1, \\ 0.8^* S_{sl} + 0.2^* S_t, & S_{sl} > \text{Threshold}, \\ 0.2^* S_{sl} + 0.8^* S_t, & S_{sl} <= \text{Threshold}. \end{cases} \tag{3}$$

where S_s, S_l, and S_t is, respectively, the syntactic, linguistic, and taxonomy-based similarity of s_1 and s_2; S_{sl} is the average of S_s and S_l; and Threshold is a given parameter to filter the matching pairs with low similarity.

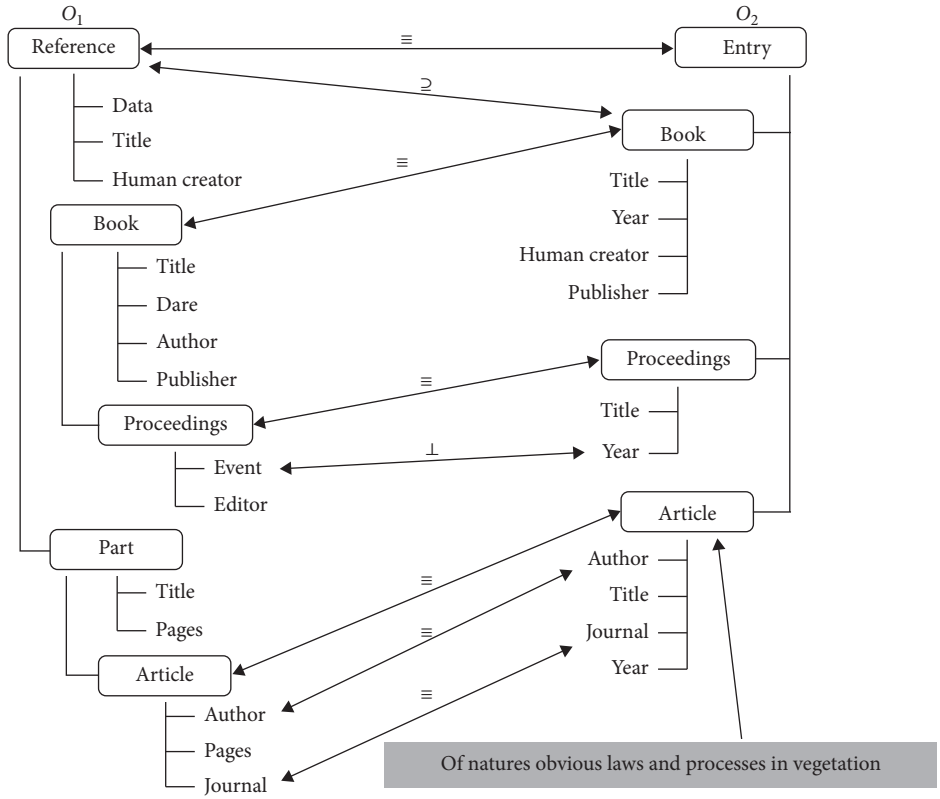


Figure 2: An example of matching two ontologies.

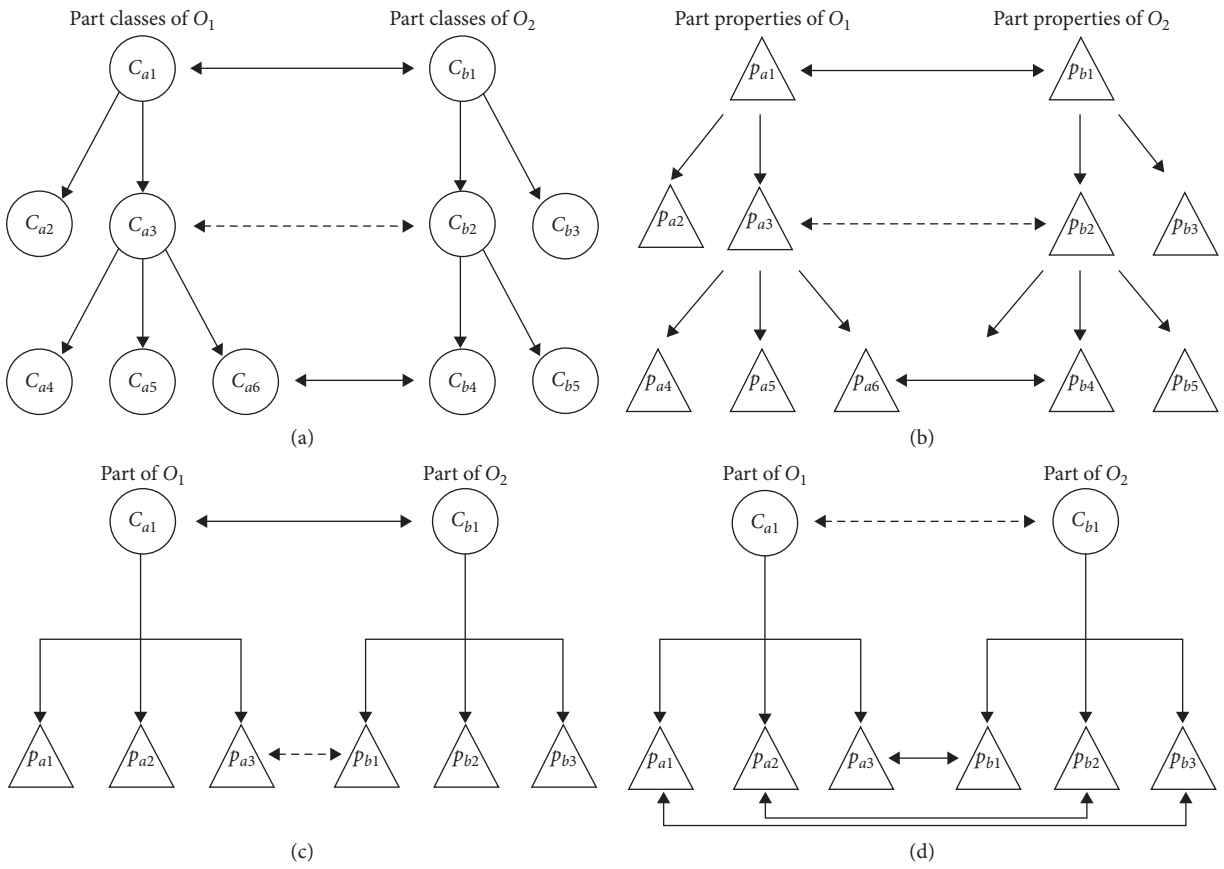


Figure 3: MRCP taxonomy-based measure.

4. Extended Compact Genetic Algorithm-Based Ontology Entity Matching

GA is an excellent methodology to solve the ontology matching problem due to its potential parallel search characteristic and good searchability. In our work, an ECGA to efficiently address the ontology entity matching problem is proposed.

4.1. Optimal Model. The optimal model of ontology entity matching problem is given as follows:

$$\begin{cases} \max & \xi(\sigma), \\ \text{s.t.} & \sigma = \left(x_1, x_2, \ldots, x_{|O_1|}\right)^T, \\ & x_i \in \left\{1, 2, \ldots, |O_2|, -1\right\}, \quad i = 1, 2, \ldots, |O_1|, \end{cases} \tag{4}$$

where $|O_1|$ and $|O_2|$, respectively, are the cardinalities of two ontologies O_1 and O_2; x_i, $i = 1, 2, \ldots, |O_1|$ is the i_{th} matching pairs. Particularly, it means there is no matching of i_{th} entity in O_1 when $x_i = -1$. The objective of this work is to maximize $\xi(\sigma)$, and for the details of it, refer to Section 4.3.3.

4.2. The Framework of ECGA-OEM. Two ontologies are to be matched as input and a reference alignment as output, and the framework of ECGA-OEM is shown in Figure 4, whose critical components are narrated in the rest of this section.Two ontologies, generally in XML or RDF format, are extracted into two hierarchy schema in the preprocessing model. The operation of the ECGA optimization model relies on the similarity matrix obtained in the similarity measure model, which has been stated in detail in Section 3.2. Finally, the alignment is generated by the solution generation model. In detail, the ECGA optimization model is described in Section 4.3.

4.3. ECGA Optimization Model. Given the virtual population's maximum generation, MaxGeneration = 2000 (normally, the number of iterations in the convergence of ECGA is much less than MaxGeneration), Threshold = 0.7, and the hierarchy schema of ontology1 and ontology2 as input and final alignment as output. The pseudocode of ECGA is proposed in Algorithm 1, where PV and BB are probability vector (see also Section 4.3.1) and building blocks (refer to also Section 4.3.7), respectively.

4.3.1. Probability Vector Initialization. Different from binary coding, the probability vector (PV) in this work is two-dimensional. The initialized PV and convergent PV are shown in Tables 1 and 2, respectively. The value in the i_{th} row and j_{th} column represents the possibility of matching between the i_{th} entity in O_1 and the j_{th} entity in O_2; i.e., it means the probability of 0_{th} entity of O_1 (reference) and 0_{th} entity of O_2 (entry) is 1/1 + 1 + 1 + 1 + 1, which is shown in Table 1 (peculiarly, the header "−1 (null)" denotes the probability of no matching). The convergence condition is that the probability of taking a unique number on each locus in the PV is 1; i.e., in Table 2, the probability of the 0_{th} entity of O_1 (reference) and the 0_{th} entity of O_2 (entry) is 23.72/0 + 23.72 + 0 + 0 + 0 and so do the rest.

4.3.2. Chromosomes Generation. Certain size chromosomes are produced in each generation through PV. An example of chromosome is given in Figure 5. In particular, subgraph (a) shows the locus of the chromosome and corresponding code, and subgraph (b) illustrates decoding chromosome in subgraph (b); i.e., it denotes that the fourth entity of O_1 "Article" correspondent to the third entity of O_2 "Article" as the code of the fourth locus is "3" ("−1 (Null)" indicates no matching).

4.3.3. Fitness Function. The fitness function is used to determine which chromosomes in the population can better adapt the environment. In the context of ontology matching, the objective of fitness function is to find the best chromosome, whose corresponding alignment's quality is the highest, with algorithm convergence. The objective function of the optimal model is used as the fitness function of this work, and given a chromosome σ, its fitness function is defined as follows:

$$\xi(\sigma) = 2 \cdot (\beta \cdot \phi(|A|) + (1 - \beta) \cdot f(A)), \tag{5}$$

where $\xi(\sigma)$ is the fitness function that is used in this paper; A is the alignment determined by σ; $|A|$ is the cardinality of A; β, a fraction in the range [0, 1], is the relative weight of $\phi(|A|)$ and $f(A)$, which is set to 0.25 in this paper; ϕ is a normalization function; and f is a function that calculates the mean of the matched entity pairs' similarity values in A. In addition, $\phi(|A|)$ and $f(A)$ are defined as follows:

$$\phi(|A|) = \frac{|A|}{\min\left(|O_1|, |O_2|\right)}, \tag{6}$$

$$f(A) = \frac{\sum_{i=1}^{|A|} \eta_i}{|A|}, \tag{7}$$

where $|O_1|$ and $|O_2|$ are, respectively, the cardinality of O_1 and O_2 and η_i is the similarity of the i_{th} matching pair in alignment A. In particular, $\phi(|A|)$ is the ratio of the number of matching pair found to the value of the smaller entity number of the two ontologies and $f(A)$ is the average similarity of the matching pairs found, which, respectively, approximates the recall value and precision value.

4.3.4. Selection Operator. The selection operator selects the best chromosomes in the current population to participate in the next step [41] and updates the PV. Firstly, the chromosomes are sorted in descending order according to their fitness scores. Secondly, the first half of the chromosomes will be retained as a temporary population. Finally, with fitness as the weight through roulette, we can select the chromosome from the temporary population for subsequent operation.

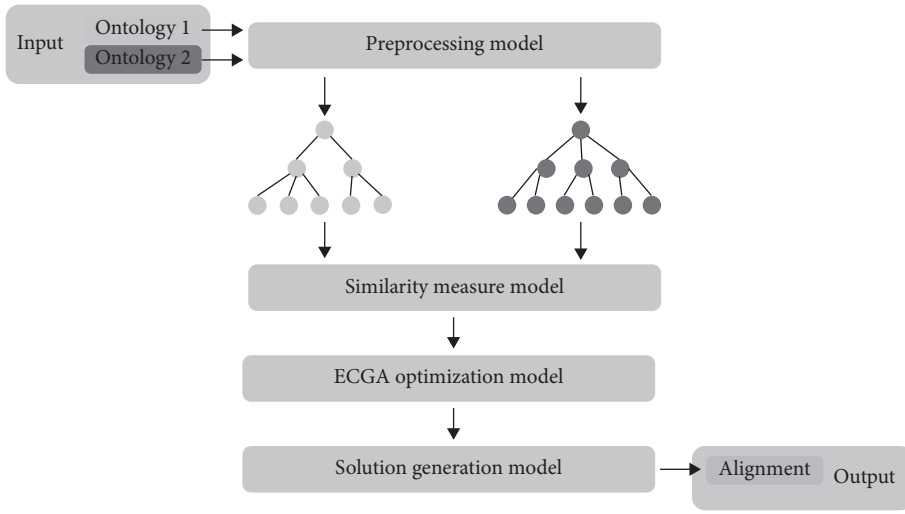


Figure 4: The framework of ECGA-OEM.

Input: the hierarchy schemas of O_1 and O_2; the aggregated similarity matrix, M_{as};
Output: the best chromosome, $\text{Best}_{\text{chromosome}}$;
(1) PV = initialization ($\text{len}_{of}O_1$, $\text{len}_{of}O_2$)
(2) **While**CurrentGeneration < MaxGeneration**do**
(3) chromosomes = generation (PV, BB);
(4) Elite = compete (Elite, chromosomes);
(5) Betters = select (Elite, chromosomes);
(6) PV = update (PV, Betters);
(7) BB = LinkageLearning (BB, PV);
(8) **end while**

Algorithm 1: ECGA optimization model.

Table 1: An example of initialized PV.

PV	0 (Entry)	1 (Book)	2 (Proceedings)	3 (Article)	−1 (Null)
0 (Reference)	1	1	1	1	1
1 (Book)	1	1	1	1	1
2 (Proceedings)	1	1	1	1	1
3 (Part)	1	1	1	1	1
4 (Article)	1	1	1	1	1

Table 2: An example of convergent PV.

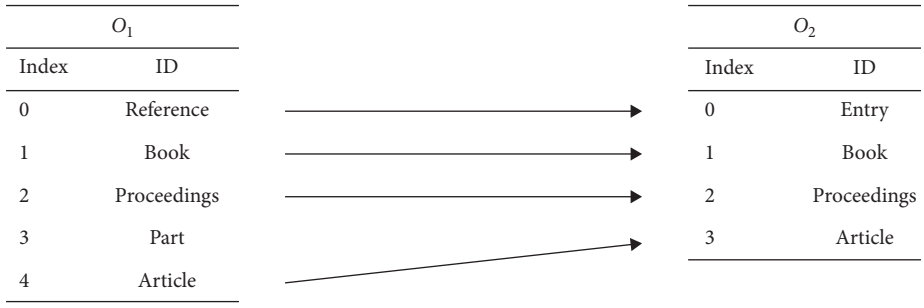
PV	0 (Entry)	1 (Book)	2 (Proceedings)	3 (Article)	−1 (Null)
0 (Reference)	23.72	0	0	0	0
1 (Book)	0	25.33	0	0	0
2 (Proceedings)	0	0	26.01	0	0
3 (Part)	0	0	0	0	22.46
4 (Article)	0	0	0	24.11	0

4.3.5. Elite Strategy. The goal of elite strategy is to keep the historical optimal solution and prevent the fitness of the optimal chromosome from "degenerating" in the process of evolution. In this work, the elite strategy has two steps: the historical optimal solution Elite first competes with the current optimal solution Best, and the winner will become the new Elite; the historical optimal solution then participates in the PV update in each generation.

4.3.6. Probability Vector Updating. An example of updating the PV is presented in Figure 6, where subgraph (a) is a chromosome generated by the initialized PV. The similarity is derived from the similarity matrix according to the chromosome's code. It should be noted that the code of the third locus is "0," which means that the entity "Part" with sequence number "3" in O_1 does not match any entity in O_2, so its similarity is equal to 1 minus the value of the highest

Index	0	1	2	3	4
Code	0	1	2	−1	3

(a)



(b)

Figure 5: An example of chromosome.

Index	0	1	2	3	4
Code	0	1	2	−1	3
Similarity	0.92	1.00	1.00	0.62	1.00
Normalized similarity	0.20	0.22	0.22	0.14	0.22

(a)

PV	0 (Entry)	1 (Book)	2 (Proceedings)	3 (Article)	−1 (Null)
0 (Reference)	1.2	1	1	1	1
1 (Book)	1	1.22	1	1	1
2 (Proceedings)	1	1	1.22	1	1
3 (Part)	1	1	1	1	1.14
4 (Article)	1	1	1	1.22	1

(b)

Figure 6: An example of update PV.

similarity between "Part" and all entities in O_2, i.e., $(1 - 0.38) = 0.62$. PV will be updated by the normalized similarity of each locus, which is shown in subgraph (b). The probability of being updated in the PV is bold; i.e., the probability of matching pair "Reference" and "Entry" changed from "1" to "1.2" since the normalized similarity of the corresponding locus of chromosome is "0.20."

4.3.7. Linkage Learning. Building blocks will be saved through linkage learning, thus to improve the efficiency of algorithm and quality of solution [42]. In simple GA, linkage learning is to identify great locus and protect them so that they will not be destroyed in the subsequent crossover and mutation operations; in ECGA, linkage learning keeps good probability distribution so that they are not disturbed in the subsequent update process. A linkage learning approach is proposed in this work, and its detail is shown in Figure 7.

For clarity, only column 1 of the original probability vector is selected for narration. In each generation, a low probability clearing operation is performed. The value of each column is divided by the maximum value of the row, and the value of the column will be cleared if the decimal is less than a specific value (0.2 in this work); i.e., 0, 2, 3, and −1 bits in PV are cleared and marked. The row of PV is a "good" probability distribution when all but one bit of this row are zero. Link learning generates building blocks based on the "good" probability distribution, i.e., a building block, the pair of index "0," and code "0," which was included in the rounded rectangle produced by linkage learning. After that, all the building blocks are directly put into each chromosome (the bold numbers), which reduces the consumption of runtime and memory consumption.

5. Experiments and Discussion

5.1. Experiment Setup. In the experiment, the Biblio benchmark provided by the Ontology Alignment Evaluation Initiative (OAEI) is used to verify the effectiveness of our approach. Normally, two ontologies to be matched and a reference alignment are included in each testing case as a standard to evaluate the quality of matching results. The testing cases can be classified into five categories, which are briefly described in Table 3.

In this work, the method is compared with the participants of OAEI-, GA-, and CGA-based ontology matching techniques. The experimental results are the H-mean values of 30 independent runs.

PV	0 (Entry)	1 (Book)	2 (Proceedings)	3 (Article)	−1 (Null)
1 (Book)	1.25	25.33	1.36	3.87	2.14

Low probability clearing

PV	0 (Entry)	1 (Book)	2 (Proceedings)	3 (Article)	−1 (Null)
1 (Book)	0	25.33	0	0	0

Linkage learning

Index 1 Code 1 Building block

Chromosome generation

Index	0	1	2	3	4
Code	0	1	2	−1	3

Figure 7: An example of linkage learning.

Table 3: A brief description of OAEI Biblio benchmark.

Cases	Brief description
101–104	The ontologies to be matched are the same or there is only a slight difference in OWL restriction between them
201–210	The ontologies to be matched have the same structural, but different lexical and linguistic features
221–247	The ontologies to be matched have the same lexical and linguistic, but different structural features
248–266	The ontologies to be matched have different lexical, linguistic, and structure features
301–304	The ontologies in the real world often have strong heterogeneity

5.2. Alignment Evaluation Metrics according to Reference Alignment. A criterion is needed when evaluating the quality of matching systems. Given an alignment result A, two measures, recall and precision, are employed in this work and their formula is as follows [22, 23]:

$$\text{recall}(A) = \frac{|R \cap A|}{|R|}, \tag{8}$$

$$\text{precision}(A) = \frac{|R \cap A|}{|A|}, \tag{9}$$

where $|R|$ and $|A|$ are the cardinality of matching pairs in the reference alignment provided in case set and the matching pairs in alignment are produced by the matching system and $|R \cap A|$ is the cardinality of matching pairs, which exist in both reference alignment and alignment found. It means that all the matching pairs in reference alignment have been found when recall is 1, while all the matching pairs found is correct when precision is 1.

Both recall and precision are important parameters of the evaluation results and they should be considered at the same time. A weighted harmonic mean of recall and precision, F-measure, is used in this work, which is presented in the following equation[43]:

$$f - \text{measure}(A) = \frac{\text{recall}(A) \cdot \text{precison}(A)}{\alpha \cdot \text{recall}(A) + (1-\alpha) \cdot \text{precison}(A)}, \tag{10}$$

where $\alpha[0, 1]$ is the relative weight of recall and precision and it is set as 0.5 in this work, which is named $f1$-measure.

5.3. Comparison with OAEI's Participants. The participants from OAEI 2016, 2015, and 2014 are selected to compare with our approach. In particular, if a matching system has participated for more than one year, only the latest results are used. The harmonic mean comparison of participants and ECGA-OEM is shown in Figure 8. The vertical axis represents different matching systems, and the horizontal axis represents the score of their corresponding parameters. In terms of f-measure, ECGA-OEM ranks third, which it is slightly lower than Lily and CroMatch. Wiki has been used as the linguistic measure in Lily and CroMatch, which is able to improve the performance of the algorithm at the expense of efficiency. In addition, ECGA-OEM has an unparalleled performance in maintaining the balance between precision and recall, while one of them is much higher than the other in the participants, which is very important in evaluating results quality.

Further f-measure comparison of OAEI participants and ECGA-OEM in a total of 32 test cases is given. Figure 9 shows the numbers of participants with ECGA-OEM superior, equal, and inferior, respectively. The horizontal axis is the set of different test cases, and the vertical axis is the number of matching systems. In the vast majority of test cases, the number of matching systems with ECGA-OEM superior to is much higher than those ones with ECGA-OEM inferior to. Only in No. 246, No. 247, and No. 254 cases, the ranking of ECGA-OEM is relatively low (refer to also Table 4 for the specific f-measure values in each testing case).

In Table 4, the numbers from 1 to 18 in the first row are edna, AML, CroMatch, Lily, LogMap, LogMapLt, Xmap,

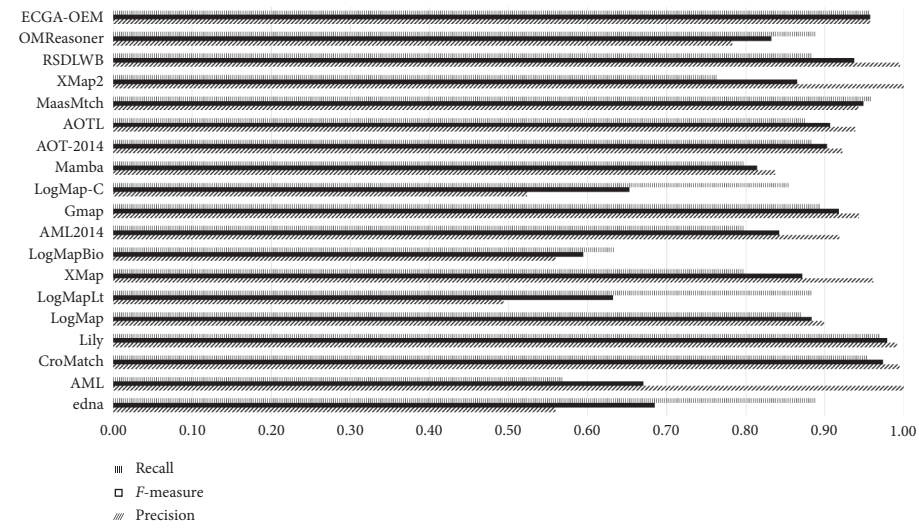


Figure 8: The harmonic mean of OAEI's participants and ECGA-OEM.

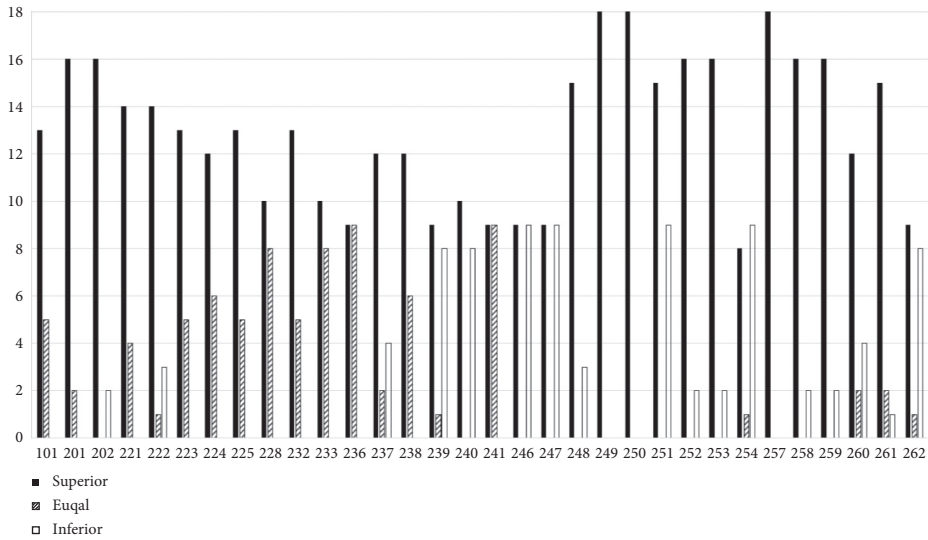


Figure 9: The number of participants with ECGA-OEM superior, equal, and inferior.

LogMapBio, AML-2014, Gmap, LogMap-C, Mamba, AOT-2014, AOTL, MassMatch, OMReasoner, RSDLWB, and Xmap2, respectively. ECGA-OEM is the matching system proposed by us. The value of each column represents the *f*-measure score of the matching system in the corresponding case. The *f*-measure of participants higher than that of ECGA-OEM is bold, and the equal ones are underlined. The "+," "=," and "−" in the last column, respectively, indicate the number of matching systems, with ECGA-OEM superior, equal, and inferior.

5.4. Comparison among GA, CGA, and ECGA. To verify the performance of linkage learning, we compare ECGA with GA and CGA. The detailed f-measure and runtime of the three competitors are, respectively, shown in Tables 5 and 6. All the GA, CGA, and ECGA's results are the mean value of 30 independent runs. It can be seen that the replacement of crossover and mutation operators (GA) with probability vector (CGA) improves the *f*-measure and significantly reduces the runtime. The average f-measure is slightly improved, while the average runtime is reduced from 31.975 seconds to 3.540 seconds; i.e., it largely improves the algorithm's efficiency with only takes about 1/10 of runtime. Except testing cases No. 301 and No. 304, CGA is more stable than GA in terms of both *f*-measure and runtime since their standard deviation is smaller. Testing cases No. 301 and No. 304 are the representatives of real-world cases with unique heterogeneity, which make the *f*-measure produced by CGA decreased slightly. Linkage learning, the technique applied in ECGA, further increased the score of *f*-measure and made decrement in runtime with average 1.749 seconds based CGA. A smaller standard deviation than CGA was obtained by ECGA, which certified the strong stability of ECGA. It is worth to be noticed that the *f*-measure score of ECGA in testing case No. 301 and No. 304 is almost the same as that of GA (only 0.003 score lower in testing case No.

Table 4: Comparison on *F*-measure harmonic mean on each testing case among OAEI participants and ECGA-OEM.

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	ECGA-OEM	+/=/-
101	0.97	0.96	1	1	1	0.93	0	0.98	0.73	0.87	0.78	0	1	1	0.95	0.71	0.97	0.52	1	13/5/0
201	0.95	0.94	0.93	0.61	0.89	0.82	0.68	0.95	0.62	0.77	0.62	0.44	1	1	0.84	0.62	0.77	0.47	1	16/2/0
202	0.81	0.84	0.92	0.81	0.86	0.82	0.64	0.87	0.62	0.76	0.62	0.42	0.98	0.99	0.85	0.62	0.76	0.48	0.96	16/0/2
221	0.97	0.95	1	1	0.98	0.93	0.99	0.98	0.73	0.87	0.78	0.51	1	1	0.94	0.72	0.97	0.53	1	14/4/0
222	0.97	0.95	1	0.8	0.98	0.94	0.77	0.99	0	0.85	0.77	0.5	1	1	0	0.72	0.78	0	0.99	14/1/3
223	0.97	0.96	1	1	1	0.93	0.99	0.98	0.73	0.86	0.78	0.51	1	1	0.94	0.72	0.97	0.53	1	13/5/0
224	0.96	0.95	1	1	1	0.93	0.97	0.98	0.92	0.87	1	0.51	1	1	0.94	0.9	0.97	0.53	1	12/6/0
225	0.97	0.96	1	1	1	0.93	0.99	0.98	0.73	0.87	0.78	0.51	1	1	0.95	0.72	0.97	0.52	1	13/5/0
228	1	1	0.93	1	1	0.84	0.97	0.96	0.5	0.88	0.55	1	1	1	0.92	0.48	1	0.8	1	10/8/0
232	0.97	0.96	1	1	0.98	0.93	0.97	0.98	0.93	0.87	1	0.51	1	1	0.94	0.9	0.97	0.53	1	13/5/0
233	1	1	0.93	1	1	0.84	0.97	0.96	0.5	0.88	0.55	1	1	1	0.92	0.48	1	0.8	1	10/8/0
236	1	1	0.93	1	1	0.84	0.97	0.96	0.86	0.88	1	1	1	1	0.92	0.8	1	0.8	1	9/9/0
237	0.96	0.95	1	0.8	1	0.94	0.77	0.99	0	0.85	0.99	0.5	1	1	0	0.91	0.78	0	0.99	12/2/4
238	0.97	0.95	1	1	1	0.93	0.99	0.98	0.93	0.87	1	0.51	1	1	0.95	0.9	0.97	0.52	1	12/6/0
239	1	1	0.93	1	1	0.84	0.97	0.96	0.5	0.88	0.55	1	1	1	0.92	0.48	1	0.8	0.97	9/1/8
240	1	1	0.93	1	1	0.84	0.97	0.96	0.5	0.88	0.55	1	1	1	0.92	0.48	1	0.8	0.99	10/0/8
241	1	1	0.93	1	1	0.84	0.97	0.96	0.86	0.88	1	1	1	1	0.92	0.8	1	0.8	1	9/9/0
246	1	1	0.93	1	1	0.84	0.97	0.96	0.86	0.88	1	1	1	1	0.92	0.8	1	0.8	0.98	9/0/9
247	1	1	0.93	1	1	0.84	0.97	0.96	0.86	0.88	1	1	1	1	0.92	0.8	1	0.8	0.99	9/0/9
248	0.83	0.84	0.93	0.81	0.86	0.82	0.84	0.87	0.63	0.76	0.62	0.45	0.98	0.97	0.8	0.62	0.77	0.43	0.9	15/0/3
249	0.81	0.8	0.87	0.65	0.89	0.82	0.81	0.87	0.81	0.76	0.79	0.42	0.95	0.95	0.82	0.8	0.82	0.48	0.96	18/0/0
250	0.81	0.83	0.85	0.88	0.88	0.72	0.87	0.84	0.42	0.76	0.43	0.9	0.94	0.98	0.81	0.4	0.88	0.69	1	18/0/0
251	0.83	0.82	0.94	0.6	0.87	0.83	0.59	0.87	0	0.75	0.61	0.43	0.93	0.97	0	0.61	0.59	0	0.91	15/0/3
252	0.82	0.83	0	0.86	0.89	0.82	0.78	0.87	0.61	0.75	0.62	0.45	0.97	0.95	0.83	0.62	0.76	0.49	0.92	16/0/2
253	0.83	0.83	0	0.79	0.87	0.82	0.79	0.87	0.81	0.76	0.79	0.42	0.97	0.94	0.86	0.79	0.81	0.47	0.9	16/0/2
254	0.83	0.85	0	0.88	0.88	0.72	0.89	0.84	0.42	0.76	0.43	0.92	0.94	0.98	0.81	0.4	0.88	0	0.83	8/1/9
257	0.84	0.85	0	0.88	0.88	0.72	0.87	0.84	0.74	0.76	0.79	0.92	0.9	0.95	0.88	0.72	0.88	0.74	1	18/0/0
258	0.81	0.83	0	0.6	0.88	0.83	0.58	0.87	0.68	0.73	0.79	0.45	0.93	0.95	0	0.8	0.66	0	0.91	16/0/2
259	0.83	0.85	0	0.8	0.89	0.82	0.83	0.87	0.8	0.75	0.79	0.46	0.96	0.94	0.83	0.8	0.81	0	0.92	16/0/2
260	0.81	0.82	0	0.88	0.88	0.74	0.87	0.85	0	0.79	0.43	0.9	0.93	0.95	0	0.41	0.9	0	0.88	12/2/4
261	0.8	0.85	0	0.88	0.88	0.72	0.87	0.84	0.42	0.76	0.43	0.9	0.89	0.89	0.83	0.4	0.88	0	0.89	15/2/1
262	0.81	0.83	0	0.88	0.88	0.72	0.85	0.84	0.74	0.76	0.79	0.92	0.94	0.94	0.81	0.7	0.88	0.69	0.83	9/1/8

Table 5: Comparison of *F*-measure and its standard deviation.

Testing case	GA		CGA		ECGA	
	f-Measure	St. dev	*f*-Measure	St. dev	*f*-Measure	St. dev
101	0.996	0.004	1.000	0.000	1.000	0.000
201	0.999	0.003	1.000	0.000	1.000	0.000
221	0.996	0.006	1.000	0.000	1.000	0.000
223	0.994	0.007	0.998	0.004	1.000	0.002
224	0.997	0.005	1.000	0.000	1.000	0.000
233	0.996	0.007	1.000	0.000	1.000	0.000
236	0.997	0.007	1.000	0.000	1.000	0.000
240	0.978	0.020	0.994	0.012	1.000	0.000
241	0.994	0.010	1.000	0.000	1.000	0.000
250	0.995	0.009	1.000	0.000	1.000	0.000
257	0.993	0.010	1.000	0.000	1.000	0.000
301	0.606	0.003	0.586	0.025	0.603	0.007
304	0.665	0.004	0.650	0.033	0.672	0.007
Average	0.939	0.007	0.941	0.006	0.944	0.001

Table 6: Comparison of runtime and its standard deviation.

Testing case	GA		CGA		ECGA	
	Runtime (s)	St. dev	Runtime (s)	St. dev	Runtime (s)	St. dev
101	44.296	4.447	4.146	0.255	2.016	0.047
201	43.463	6.195	4.144	0.279	1.976	0.042
221	43.204	5.365	4.115	0.223	2.038	0.041
223	51.146	6.640	6.947	0.208	3.226	0.255
224	44.624	6.106	4.143	0.277	2.053	0.058
233	12.686	3.080	1.276	0.078	0.734	0.022
236	13.463	2.797	1.244	0.054	0.720	0.030
240	22.740	4.950	3.294	0.028	1.771	0.145
241	13.474	3.804	1.236	0.076	0.728	0.024
250	14.256	2.576	1.235	0.093	0.624	0.026
257	14.507	2.765	1.264	0.082	0.611	0.028
301	45.056	4.253	6.306	0.880	3.353	0.086
304	52.759	6.400	6.673	0.705	2.885	0.067
Average	31.975	4.568	3.540	0.249	1.749	0.067

301 and this minor difference could be ignored), which embodied the ability of linkage learning to overcome the CGA's shortcoming.

6. Conclusions

Ontology matching can effectively solve the problem of data heterogeneity by discovering correspondence between two ontologies' entities. Compact encoding mechanism shows high efficiency in ontology matching, especially in ontology entity matching. Linkage learning, which is employed in ECGA-OEM proposed in this paper, can produce qualified alignment of ontology matching. The experiment results have shown that our approach outperforms the participants of OAEI in terms of f-measure, recall, and precision. The comparison on terms of f-measure and runtime with GA, CGA, and ECGA shows that ECGA-OEM is able to greatly reduce the runtime consumption while maintaining the alignment's quality.

In the future, we would like to apply the linkage learning in other compact SIAs for better results. In addition, matching large-scale ontologies, such as anatomy and large biomedical track in OAEI, are an open challenge in the domain of ontology matching. We are interested in using the improved ECGA to match these large-scale ontologies.

Data Availability

The data used to support the findings of this study are available upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province (no. 2020J01875) and the National Natural Science Foundation of China (nos. 61773415, 61801527, and 61103143).

References

[1] X. Xue, J. Liu, P.-W. Tsai, X. Zhan, and A. Ren, "Optimizing ontology alignment by using compact genetic algorithm," in *Proceedings of the 2015 11th International Conference on Computational Intelligence and Security (CIS)*, pp. 231–234, IEEE, Shenzhen, China, December 2015.

[2] X. Xue, "A compact firefly algorithm for matching biomedical ontologies," *Knowledge and Information Systems*, vol. 62, no. 11, pp. 1–17, 2020.

[3] J. M. V. Naya, M. M. Romero, J. P. Loureiro, C. R. Munteanu, and A. Pazos Sierra, "Improving ontology alignment through genetic algorithms," in *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies*, pp. 240–259, IGI Global, Hershey, PA, USA, 2010.

[4] J. Bock and H. Jan, "Discrete particle swarm optimisation for ontology alignment," *Information Sciences*, vol. 192, pp. 152–173, 2012.

[5] He Yao, X. Xue, and S. Zhang, "Using artificial bee colony algorithm for optimizing ontology alignment," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 4, pp. 766–773, 2017.

[6] J.-S. Pan, P.-C. Song, S.-C. Chu, and Y.-J. Peng, "Improved compact cuckoo search algorithm applied to location of drone logistics hub," *Mathematics*, vol. 8, no. 3, p. 333, 2020.

[7] A.-Q. Tian, S.-C. Chu, J.-S. Pan, H. Cui, and W.-M. Zheng, "A compact pigeon-inspired optimization for maximum short-term generation mode in cascade hydroelectric power station," *Sustainability*, vol. 12, no. 3, p. 767, 2020.

[8] N. Alboukaey and A. Joukhadar, "Ontology matching as regression problem," *Journal of Digital Information Management*, vol. 16, no. 1, 2018.

[9] M. Ali Khoudja, M. Fareh, and H. Bouarfa, "Ontology matching using neural networks: survey and analysis," in *Proceedings of the 2018 International Conference on Applied Smart Systems (ICASS)*, pp. 1–6, IEEE, Medea, Algeria, November 2018.

[10] M. T. Dhouib, C. F. Zucker, and A. G. B. Tettamanzi, "An ontology alignment approach combining word embedding and the radius measure," in *Proceedings of the 15th International Conference on Semantic Systems*, pp. 191–197, Springer, Karlsruhe, Germany, September 2019.

[11] A. Ali, H. Mcheick, K. Ahmad, and W. Dhifli, "Context-aware instance matching through graph embedding in lexical

semantic space," *Knowledge-Based Systems*, vol. 186, Article ID 104925, 2019.

[12] F. Ali, K.-S. Kwak, and Y.-G. Kim, "Opinion mining based on fuzzy domain ontology and support vector machine: a proposal to automate online review classification," *Applied Soft Computing*, vol. 47, pp. 235–250, 2016.

[13] X. Xue and J.-S. Pan, "A segment-based approach for large-scale ontology matching," *Knowledge and Information Systems*, vol. 52, no. 2, pp. 467–484, 2017.

[14] S. Amrouch, S. Mostefai, and M. Fahad, "Decision trees in automatic ontology matching," *International Journal of Metadata, Semantics and Ontologies*, vol. 11, no. 3, pp. 180–190, 2016.

[15] T. B. Araújo, C. E. Santos Pires, T. Pereira Da Nóbrega, and D. C. Nascimento, "A fine-grained load balancing technique for improving partition-parallel-based ontology matching approaches," *Knowledge-Based Systems*, vol. 111, pp. 17–26, 2016.

[16] M. B. Amin, W. A. Khan, S. Hussain et al., "Evaluating large-scale biomedical ontology matching over parallel platforms," *Iete Technical Review*, vol. 33, no. 4, pp. 415–427, 2016.

[17] X. Xue and J. Liu, "Collaborative ontology matching based on compact interactive evolutionary algorithm," *Knowledge-Based Systems*, vol. 137, pp. 94–103, 2017.

[18] X. Xue and J.-S. Pan, "A compact co-evolutionary algorithm for sensor ontology meta-matching," *Knowledge and Information Systems*, vol. 56, no. 2, pp. 335–353, 2018.

[19] X. Xue, J. Chen, J. Chen, and D. Chen, "Using compact coevolutionary algorithm for matching biomedical ontologies," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 2309587, , 2018.

[20] X. Xue and J. Chen, "A compact co-firefly algorithm for matching ontologies," in *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2629–2632, IEEE, Xiamen, China, December 2019.

[21] X. Xue and J. Chen, "Using compact evolutionary tabu search algorithm for matching sensor ontologies," *Swarm and Evolutionary Computation*, vol. 48, pp. 25–30, 2019.

[22] S.-C. Chu, X. Xue, J.-S. Pan, and X. Wu, "Optimizing ontology alignment in vector space," *Journal of Internet Technology*, vol. 21, no. 1, pp. 15–22, 2020.

[23] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, p. 2056, 2020.

[24] Y. Wang, H. Yao, L. Wan et al., "Optimizing hydrography ontology alignment through compact particle swarm optimization algorithm," in *Proceedings of the International Conference on Swarm Intelligence*, pp. 151–162, Springer, Chiang Mai, Thailand, July 2020.

[25] X. Xue and P.-W. Tsai, "Matching biomedical ontologies with compact evolutionary algorithm," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 3–10, Springer, Singapore, Singapore, May 2020.

[26] X. Xue, H. Yang, J. Zhang, J. Zhang, and D. Chen, "An automatic biomedical ontology meta-matching technique," *Journal of Network Intelligence*, vol. 4, no. 3, pp. 109–113, 2019.

[27] X. Xue and J.-S. Pan, "An overview on evolutionary algorithm based ontology matching," *ournal of Information Hiding and Multimedia Signal Processing*, vol. 9, pp. 75–88, 2018.

[28] J. Martinez-Gil, E. Alba, and J. F. Aldana-Montes, "Optimizing ontology alignments by using genetic algorithms," in *Proceedings of the Workshop on Nature Based Reasoning for the Semantic Web*, Karlsruhe, Germany, October 2008.

[29] J. Martinez-Gil and J. F. Aldana-Montes, "Evaluation of two heuristic approaches to solve the ontology meta-matching problem," *Knowledge and Information Systems*, vol. 26, no. 2, pp. 225–247, 2011.

[30] G. Alexandru-Lucian and I. Adrian, "Using a genetic algorithm for optimizing the similarity aggregation step in the process of ontology alignment," in *Proceedings of the 9th RoEduNet IEEE International Conference*, pp. 118–122, IEEE, Sibiu, Romania, June 2010.

[31] G. Acampora, V. Loia, and A. Vitiello, "Enhancing ontology alignment through a memetic aggregation of similarity measures," *Information Sciences*, vol. 250, pp. 1–20, 2013.

[32] X. Xue, Y. Wang, and A. Ren, "Optimizing ontology alignment through memetic algorithm based on partial reference alignment," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3213–3222, 2014.

[33] X. Xue and Y. Wang, "Optimizing ontology alignments through a memetic algorithm using both matchfmeasure and unanimous improvement ratio," *Artificial Intelligence*, vol. 223, pp. 65–81, 2015.

[34] J. Wang, Z. Ding, and C. Jiang, "Gaom: genetic algorithm based ontology matching," in *Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06)*, pp. 617–620, IEEE, Guangzhou, China, December 2006.

[35] G. Acampora, V. Loia, S. Salerno, and A. Vitiello, "A hybrid evolutionary approach for solving the ontology alignment problem," *International Journal of Intelligent Systems*, vol. 27, no. 3, pp. 189–216, 2012.

[36] A. Alves, K. Revoredo, and F. Baião, "Ontology alignment based on instances using hybrid genetic algorithm," in *Proceedings of the 7th International Conference on Ontology Matching-Volume 946*, pp. 242-243, November 2012.

[37] X. Xue and Y. Wang, "Using memetic algorithm for instance coreference resolution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 580–591, 2015.

[38] W. Jan Heeringa, *Measuring dialect pronunciation differences using levenshtein distance*, Ph.D thesis, University Library Groningen, Groningen, Netherlands, 2004.

[39] G. Stoilos, G. Stamou, and S. Kollias, "A string metric for ontology alignment," in *Proceedings of the International Semantic Web Conference*, pp. 624–637, Springer, Galway, Ireland, November 2005.

[40] G. A. Miller, "WordNet," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[41] D. Cai, Y. Wang, M. Ye, X. Xue, and H. Liu, "An orthogonal evolutionary algorithm with learning automata for multiobjective optimization," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 3306–3319, 2015.

[42] G. R. Harik, F. G. Lobo, and K. Sastry, *Linkage Learning via Probabilistic Modeling in the ECGA*, Springer, Berlin, Germany, 1999.

[43] X. Xue and X. Yao, "Interactive ontology matching based on partial reference alignment," *Applied Soft Computing*, vol. 72, pp. 355–370, 2018.

Hindawi
Complexity
Volume 2021, Article ID 6610797, 11 pages
https://doi.org/10.1155/2021/6610797

Research Article

A Back Propagation Neural Network-Based Method for Intelligent Decision-Making

Hao Zhang [1,2] **and Jia-Hui Mu** [3]

[1]*School of E-Business and Logistics, Beijing Technology and Business University, Beijing 100048, China*
[2]*Beijing Food Safety Research Base, Beijing, China*
[3]*School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China*

Correspondence should be addressed to Jia-Hui Mu; mu_jh123@163.com

Received 11 December 2020; Revised 28 December 2020; Accepted 4 January 2021; Published 4 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

A shortage or backlog of inventory can easily occur due to the backward forecasting method typically used, which will affect the normal flow of funds in pharmacies. This paper proposes a replenishment decision model with back propagation neural network multivariate regression analysis methods. With the regular pattern between sales and individual variables, supplemented with the safety stock empirical formula, an accurate replenishment quantity can be obtained. In the case analysis, this paper takes the sales situation of a pharmacy as an example and tests the accuracy and stability of the model. The results show that the model has good prediction accuracy which can be introduced into the intelligent pharmacy system and used in the replenishment of the intelligent pharmacy to prevent overstocking or a shortage of stock, thus improving the financial situation, reducing the manpower burden of typical retail pharmacy, and helping residents buy medicines.

1. Introduction

The pharmacy business model has gradually entered the public view. It refers to the upgrading of the pharmaceutical industry and retail terminals from traditional business models to innovative industrial models through innovative technologies such as the Internet, big data, and artificial intelligence. Compared with traditional pharmacies, the intelligence of pharmacies is mainly reflected in the client, enabling customers to purchase medicines more conveniently. However, the method for forecasting the amount of resupply for a pharmacy is still backward. Many companies estimate safety stocks based on experience (Trapero et al.) or simply use software provided by software vendors to make predictions for decades without updating the system [1]. The data provided by software vendors are too conservative which has resulted in low turnover of goods and slow financial flows [2]. To solve this problem, two aspects should be considered. The first is how to determine when and how much to replenish, and the second is how to predict demand.

Many scholars have expressed their opinions on the first aspect. The classic single models are the moving average (Holt [3]), index smoothing method, and linear regression method (Galton, 1855), [4, 5]. Because these methods have limitations, many new models have emerged, such as the grey model [6, 7], wavelet smoothing model [8], autoregressive model (Arima) [9], and neural network [10–12]. However, these methods have their own limitations. The grey model when the absolute value of the development coefficient is large ($-a > 0.5$), the model deviation is large, and it cannot be used for medium- and long-term predictions, and even short-term predictions are not appropriate. The wavelet smoothing model can flexibly adapt to sudden data changes but performs very poorly in smooth time series prediction. The autoregressive model is effective. However, this method also has its own applicability; that is, these data must be autocorrelated, and the autocorrelation coefficient must be greater than 0.5; otherwise, the prediction result will be extremely inaccurate. The neural network model also has its applicability most

notably because the model is input and then learned by each data when learning, which causes it to be easily confined to a small branch of the entire data [13].

After discovering that a single model has its own defects and applicability, many scholars began to study the combination model to reduce the prediction error. Wang et al. found that combination forecasting is one of the most important and effective methods in time series forecasting, so they proposed a combination forecasting model based on the neural network linear integration framework and compared other four single neural network models and other forecasting models. The results show that the performance of NNSLEF is better than that of the four component neural network model and other recognized models [14]. Choi et al. proposed a hybrid prediction model. By combining the classic SARIMA method and the wavelet transform model, they were able to predict the sales data and test it against real data to prove its effectiveness [15]. Christodoulos et al. with limited historical data and short-term projections of the penetration rate of broadband and mobile communications in the world, combined with ARIMA and the diffusion model, proved the accuracy of their method when applied to broadband and mobile phone penetration worldwide [16]. Zhang et al. proposed a novel solution by using common spatial pattern (CSP) and convolutional neural network (CNN) on disease prediction, and the experimental result demonstrates that the proposed approach outperforms most methods [17]. Wang et al. proposed a new monthly tourism demand forecasting method by combining improved chaotic particle swarm optimization (ICPSO) with backpropagation neural network (BPNN) to forecast monthly tourism demand. The results show that the performance of the ICPSO-BPNN model is better than that of the basic BPNN model, autoregressive comprehensive moving average model, support vector regression model, and other popular models [18].

Another point of concern is to determine the safety of stocks, the replenishment points, and the spacing of replenishment. Cui et al. and Foster et al. studied the tendency of goods to be out of stock with customers still wanting to buy them and elaborated on the impact of the lack of goods on the purchases of customers in different situations [19, 20]. When making replenishment decisions, Puga et al. found that the two-stage supply chain model, which uses a service-level integration (s, Q) supply strategy, was the best and could improve inventory turnover [21]. Trapero et al. used the GARCH model to calculate the safe inventory and compared it with the kernel density estimation and found that the kernel density estimation GARCH model is suitable for shorter periods of time, while the GARCH model is more suitable for longer periods of time [1]. Predicted intervals are generally calculated using empirical algorithms, unless it is not a fixed replenishment interval, but some scholars used the neural network approach instead [22].

Although some commonly used methods and scholarly research results related to drug replenishment currently exist, these methods have some limitations, such as the index smooth method, which is applicable only for short-term measurements. Applying this method to long-term predictions can easily yield inaccurate results. The wavelet smoothing method is very sensitive to data fluctuations, but the prediction error for stable data is very large [8]. Although the grey system analysis model looks very precise on the surface, the error should be small, and the prediction time must be very accurate. However, when the value of "a" (the basic relations of the grey system model is as follows: $a^{(1)}(x^{(1)}(k+1)) = -aX^{(1)}(K+1) + u$) is large, the error will become large, and large deviations in the prediction will occur [7]. To address the limitations of these methods, this paper combines the back propagation neural network and regression analysis in the demand forecasting stage and brings together the highly similar characteristics of neural networks and the sensitivity of multiple regression to seasonal changes to better predict the demand for future drugs. Moreover, this method has not yet been studied by scholars in drug replenishment; hence, the research space is very large. It is suitable for short-term and medium-term predictions and can meet the replenishment of commonly used drugs, especially for seasonally sensitive drugs. For safety stock and replenishment points, this paper uses the traditional empirical formula, which is simple but can effectively calculate the location of safety stock and replenishment points.

The rest of the contents are organized as follows. Section 2 introduces the model and methods used in the paper. In Section 3, the framework of the replenishment decision model is introduced, and an example of cold medicine in a pharmacy is taken to examine the model validity. Section 4 shows the result of the replenishment decision and analyses the predictive effect of the methods. Finally, the conclusions are given in Section 5.

2. Model Background

2.1. Replenishment Decision Model. The empirical formula helps to build up the replenishment decision model and is commonly used to calculate the safety stock, replenishment points, prediction interval, and replenishment quantity. Because medicines are stored in obsolete stocks, excessive stocks can lead to a backlog of stocks, a low turnover of funds, and a tendency to strain financial flows, while the scarcity of stocks, increasing inventory turnover, may lead to stock-outs. Therefore, a proper inventory is essential for the operation of pharmacies. Hence, this paper must use the formula to accurately fill in the amount of goods to achieve the maximum benefit.

Here, the paper assumes that the spacing of the replenishment is not fixed and does not consider the cost of transporting drugs because each drug delivery involves several orders carried together, the variety of drugs is large, and the cost of sharing each package is negligible. At the same time, the inventory cost of storing drugs in pharmacies is not considered here. Since the paper is based on the sales quantity and stock limit to set the amount of replenishment, the shelf life of drugs will not be exceeded. The paper considers only the point of replenishment, the safety stock,

the upper and lower levels of the inventory, and the predicted demand. Finally, the replenishment quantity is determined based on these main factors.

First, the symbols needed in the paper are defined in Table 1:

The safety stock empirical formula is as follows:

$$\mathrm{SC} = R * \sigma * \sqrt{T}. \tag{1}$$

The formula for calculating the replenishment point is as follows:

$$D = V_T + \mathrm{SC}. \tag{2}$$

When the pharmacy inventory is less than this point, the restocking action begins.

$$V_K = S_K. \tag{3}$$

This paper assumes that the turnover rate of the inventory is not fixed, and the predicted interval is the time at which the inventory is roughly turned over once. Here, we set the forecast period as the average time for drug inventory turnover and assume that the demand is approximately equal to the sales forecast for a certain range in the future.

$$V = \begin{cases} V_K, & V_K + \mathrm{SC} < V_{\max}, \\ V_{\max}, & V_K + \mathrm{SC} \geq V_{\max}. \end{cases} \tag{4}$$

If replenishment is required, the demand for replenishment intervals is compared with the safety stock ceiling. If the former is smaller than the latter, the replenishment quantity is equal to the demand for replenishment intervals; otherwise, it is the upper limit of the inventory. Replenishment can be carried out according to this premeasurement. Then, the drug replenishment forecast concludes.

2.2. Multiple Regression Analysis. Multiple regression analysis refers to the selection of one variable as the dependent variable and other variables as independent variables among all the relevant variables. Then, according to these variables, a linear or nonlinear mathematical expression among several variables is established, and the sample data are used to test and analyse the statistical method. The expression for the multiple regression analysis is as follows:

$$Y_i = m_0 + m_1 X_1 + m_2 X_2 + \cdots + m_k X_k + e, \tag{5}$$

where Y is the dependent variable, X_k is the independent variable, m_0 is the constant term, m_1, m_2, …, m_k are the regression coefficients, and e is the error, which follows the normal distribution with a mean value of zero. The regression coefficient can be solved by the least square method as $m = (X^T X)^{-1} X^T Y$. The general form of the multiple regression analysis method is as follows:

$$Y_i = m_0 + m_1 X_{1i} + m_2 X_{2i} + \cdots + m_k X_{ki}, \tag{6}$$

Table 1: Symbol definitions for the replenishment decision model.

Symbol	Definition
SC	Safety stock
R	Safety factor
σ	Standard deviation of demand
T	Lead time (days)
D	Order point
V	Replenishment quantity
V_T	Lead time requirements
$V_{\max}$	Maximum stock quantity
$V_{\min}$	Minimum stock quantity
K	Prediction interval
V_K	Demand of prediction interval
S	Sales forecast
S_k	Sales in prediction interval

where $i = 1, 2, 3, \ldots, n$. After variable replacement, the multiple regression analysis method can be converted into a multiple linear regression method and solved.

2.3. Back Propagation Neural Network. The back propagation neural network (BPNN) is one type of ANN. It was proposed in 1986 by a team of scientists, including McCelland and Rumelhart, and is one of the most commonly used neural network methods. The method is a multilayer feed-forward neural network trained by the error back propagation algorithm. The basic principle is to learn the input sample, judge the error, modify the weight and the threshold value to reduce the error, and then iterate many times to obtain the optimal mapping relationship.

The BPNN is composed of many layers, including the input layer, hidden layer, and output layer. The structure of the BPNN prediction method is shown in Figure 1. In the figure, x_n is the input layer, which shows the time series data of each index, z_q is the hidden layer, and y_m is the output layer, which shows the predicted value of each index.

The BPNN is calculated as follows:

Step 1. Each weight value is limited to an interval, and each weight value is randomly assigned within the interval.

Step 2. The error function is set, given the accuracy of the calculation and the maximum number of learning times.

Step 3. The input independent variables and corresponding dependent variables are randomly extracted.

Step 4. The input and output of each layer of neurons are calculated.

Step 5. Based on the calculated output results and the original sample dependent variables, the partial derivative of the error function relative to the output layer results is calculated.

Step 6. Using the randomly set weight, the partial derivative of the output layer, and the output result of the neural network calculation, the partial derivative of

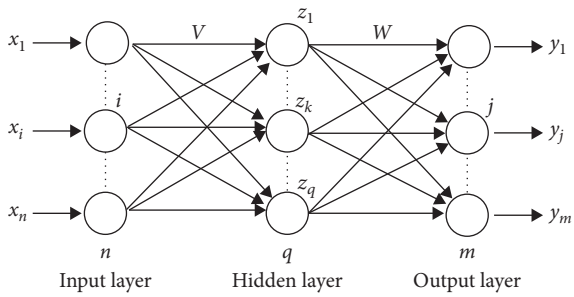


Figure 1: BPNN method.

the error function relative to the neurons of the hidden layers is calculated.

Step 7. Then, the results just calculated and the output of each neuron in the hidden layer are used to correct each weight.

Step 8. The weight value is corrected by the input value of each neuron in the input layer and the partial derivative function of each neuron in the implicit layer.

Step 9. The global error is calculated.

Step 10. Whether the global error meets the requirements is determined, and the entire process is ended when the error meets the requirements or the current number of learning times is greater than the maximum number of settings. Otherwise, the next independent variable and corresponding dependent variable are selected, and the next calculation begins.

2.4. BPNN Multivariate Regression Analysis. This paper uses BPNN multivariate regression analysis to forecast the drug demand. After regression analysis with the BPNN and multivariate regression, the method is obtained by taking the weighted combination of each method. The neural network method is Q_1, the multivariate regression analysis method is Q_2, and the weights of the neural network method and the multivariate regression analysis method are K_1 and K_2, respectively. To ensure that the results of the combination processing are unbiased, the following must be satisfied:

$$K_1 + K_2 = 1, \quad K_1, K_2 \geq 0. \tag{7}$$

The expression of its combination method is as follows:

$$Q = K_1 Q_1 + K_2 Q_2. \tag{8}$$

Set $K_1 = \mu$ and $K_2 = 1 - \mu$.

$$Q = Q_1 + (1 - \mu) Q_2. \tag{9}$$

Set ϵ to the error between the actual sales value and the predicted result value so that the actual forecast value is Q_S, then

$$\epsilon = Q_S - Q = Q_S - Q_1 - (1 - \mu) Q_2. \tag{10}$$

According to Lagrange's rule, the variance and minimum of the predicted value and the actual value can be obtained when the following formula is satisfied:

$$\sum_{i=1}^{n} \varepsilon_{\min}^2 = \sum_{i=1}^{n} \left[Q_{si} - \mu Q_{1i} - (1 - \mu) Q_{2i} \right]^2, \tag{11}$$

where i represents the number of measurements and n represents the number of measurements. Then, the final weight values of K_1 and K_2 are calculated and brought into the combination method to find its predicted values.

3. Replenishment Decision Model

This section describes in detail the decision model used to predict the amount of replenishment in pharmacies, including the data, the methods used, and the modelling process.

3.1. Establishing a Model Framework. This paper uses a model combining multiple regression analysis and the BPNN and an empirical formula to build the replenishment decision model. The detailed steps for its realization are as follows:

Step 1. Establish the replenishment decision model

The historical sales data and empirical formula are used to establish a replenishment decision model. The drug safety inventory, replenishment point, replenishment interval, and replenishment forecast formulas are included.

Step 2. Analyse and collect the data

Use the historical sales data of a drug in a pharmacy as the dependent variable of the demand forecast. Then, an analysis is conducted to identify all the factors that affect the sales of the drug as its independent variables.

Step 3. Process the data

The numerical and standardized treatment of these factors makes their regression expression neater and easier to calculate. The correlation between the data is tested. By testing the correlation between the various variables, we analyse and eliminate the correlation among the nonsignificant factors. We then observe whether a phenomenon exists in which the correlation between independent variables is greater than that between the independent variables and dependent variables. If it is, one of the factors is discarded to prevent multicollinearity.

Step 4. Forecast the replenishment

Multiple regression analysis and the BPNN model are used to find the relationship between the variables and check whether multicollinearity occurs. If not, the combined multiple regression analysis and BPNN model are used to obtain a new expression for sales and their variables. Then, the safety stock and replenishment point obtained from step 1 are employed to forecast the replenishment quantity.

The overall model structure is shown in Figure 2.

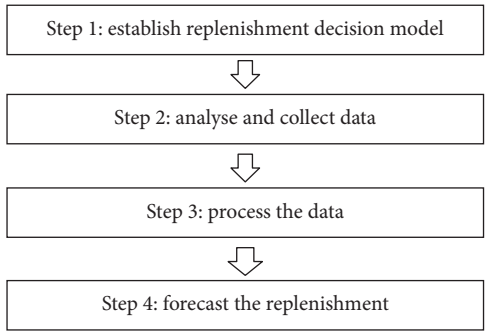


Figure 2: Model framework.

3.2. Establish the Replenishment Decision Model. To examine the effect of the decision model, a cold medicine from a pharmacy in Beijing was selected for analysis. The data used in this article were processed due to the partial design of commercial secrets, and the rest of the data are real data. The daily sales quantity of the drug from 2016 to 2018 and the factors that may affect the sales quantity were studied.

Among them, the standard differences in the demand, safety factors, and stock ceilings are all known. The sales forecast of the replenishment interval can be obtained by the multivariate regression model of the neural network established above. The remaining variables can be calculated by the known quantities.

By looking at these data, we found that a large difference in sales exists between different seasons. Therefore, this paper divides the sales time of the cold medicine into the off-season and the busy season and performs separate calculations to obtain a more accurate safety inventory that is more suitable. The off-season period is from May to October, and the busy season is from November to April.

Among these variables, the lead time $T = 3$, the level of service is 0.99, $R = 2.33$, the standard deviation of demand in the off-season is 1.46, and that in the busy season is 2.07. These values are entered into formula (1) separately to solve; the off-season safety stock is 6, and the busy season safety inventory is 8. The demand for the preseason period is obtained by multiplying the average daily sales quantity in the light season, of which the average daily sales quantity in the off-season is 3 and the average daily sales quantity in the busy season is 7. These values are entered into formula (2) separately to solve the off-season replenishment point, which is 15, and the busy season replenishment point, which is 29.

Due to the different stock turnover rates in the light season, this paper calculates the approximate turnover rates for the light season separately based on the historical data and then sets the forecast sales days for the off-season to 30 days and for the busy season to 15 days. When more than half the number of days in the forecast range are in another quarter, the paper makes up for this in another quarter. Therefore, when the stock reaches the replenishment point, this paper chooses different demand forecasts according to the light busy season, forecasts the sales quantity for the next 30 days during the off-season, and forecasts the sales quantity for the next half-month during the busy season. That is, $K = 30$ in the off-season, and $K = 15$ in the busy season.

3.3. Analysis and Collection of Data. This paper analyses all the possible factors that impact cold medicine sales (as follows) and collects the relevant data.

Incidence Factors. These factors are directly related to the number of medicines purchased. When the incidence of colds increases, more people will buy drugs. Since a cold is a "small" disease, people are more likely to go to the pharmacy to buy medicine when they have a cold.

Appropriate Travel Factors. Whether people will buy medicines or not may be affected by the appropriateness of travel. For example, when the weather is good, people may prefer to go out and buy medicine. This factor is a combination of factors, including multiple indicators such as weather, average temperature, wind power, and air pollution index.

Population Mobility Factors. When the important and long holiday season comes, people may go out to play or return home, which may lead to changes in the level of drug purchase on these days.

Promotion Factors. Although the frequency of sales promotions is not very high for drugs, the number of medicines purchased may change during a promotional period. Therefore, this factor must be explored.

3.4. Data Processing

Incidence Factors. For the incidence factors, this paper was able to obtain comparative data from 2016 to 2018 through the website of the Beijing Center for Diseases Prevention and Control. The first day of January 2016 was used as a benchmark. The difference between these data values and the baseline values determines the strength of the incidence rate. The prevalence rate is determined, and the incidence rate is expressed as 1–3, of which 1 is the weakest and 3 is the strongest.

Appropriate Travel Factors. Here, this paper took the day, weather (cloudy, sunny, rainy, and snowy), average temperature, wind, and air pollution index as factors that affect travel. The website (weather forecast 2345) was searched for daily weather data from 2016 to 2018, and these factors were quantified. The specific treatment criteria are shown in Tables 2–5.

Then, each factor has been quantified according to the data processing criteria based on the weight obtained in the questionnaire (approximately 2.3 : 1 : 1 : 1, of which the weather is 2.3 and the rest are 1). After adding the weighted data, the final suitable travel degree data are obtained.

Population Mobility Factors. Here, the paper looks for data to obtain the total population of the district covered by the pharmacy. The resident population

knows that its population is more mobile; thus, greater population mobility may occur during the holiday period. Therefore, obtain the data of legal holidays from 2016 to 2018 through the China Government Website, and the size of each value is used to indicate the mobility of the population. The specific criteria are shown in Table 6.

Promotional Factors. This thesis obtained the in-store promotion time data for 2016–2018 (to protect the privacy of the business, the data are processed here). After that, the data are in numeric form, represented by the number 2 at the time of promotion, with the remaining time represented by 1.

Sale Quantity. To determine the quantity forecast, this paper needs to predict the future demand, but the demand has an inseparable relationship with the historical sales quantity. Thus, this paper studies whether and to what extent historical sales are related to these factors and then obtains an expression of the relationship between sales quantity and these factors, after which a forecast of future demand is made. Here, the paper surveys the daily sales of a drug from 2016 to 2018 at a pharmacy and studies the 2016 and 2017 data as dependent variables (to protect the privacy of the business, the data are processed here; however, this does not affect the accuracy of the predictions).

To investigate whether independent variables exist, whether dependencies between the independent variables and dependent variables can be found, and how these variables relate to each other, the paper studies the daily sales data of a cold medicine at a pharmacy from 2016 to 2018, along with the incidence factors, appropriate travel factors, population mobility factors, and promotion factors mentioned in the previous section. A correlation test is then performed to observe the relationship .

Table 7 shows the results obtained through a one-sided test. Among them, the daily sales quantity and incidence rate, suitable travel rate, population mobility, and promotion factors are all significantly related (at the level of 0.05).

In particular, the correlation coefficient between daily sales and morbidity = 0.808 > 0.8 is highly related. The remaining daily sales are relatively weak in relation to the suitability of travel, population mobility, and promotion. Among them, daily sales are negatively related to population mobility. The correlation between independent variables is weak or not significant; thus, multi collinearity is not possible.

3.5. Forecast Replenishment. BPNN regression and multivariate regression analysis was carried out with the variables tested earlier. The results are as follows.

Because the BPNN model is prone to local overlearning during learning, the results of each test may be different. Therefore, the paper trains and analyses these data several times and takes the highest degree of similarity as the output.

Table 2: Numerical table of weather factors.

Weather conditions	Evaluation criteria
Sunny, cloudy	5
Haze, sand	4
Thundershower, light rain	3
All kinds of rain	2
Snow	1

Table 3: Numerical table of average temperatures.

Average temperature	Evaluation criteria
$[-\infty, 0)$	1
$[0, 9) \cup [29, +\infty)$	2
[9, 19)	3
[19, 29)	4

Table 4: Numerical table of wind power levels.

Power of the wind (level)	Evaluation criteria
[0, 2]	5
(2, 3]	4
(3, 4]	3
(4, 5]	2
(5, 6]	1

The degree of influence of the independent variables on the daily sales quantity is expressed in Table 8. From high to low, the order is as follows: incidence rate, promotion factor, suitable travel degree, and population mobility. Their levels of importance are 100%, 58.4%, 57.7%, and 45.7%, respectively.

Multivariate regression considers the daily data from 2016 to 2017. The regression method adopts a step-by-step approach. The following results are obtained. Table 9 reveals that the goodness of fit (R^2) is improving because the closer R^2 is to 1, the higher the degree of similarity. In the fourth model, the goodness of fit (R^2) is 0.768, indicating that this model can explain 76.8% of the variable changes and that the degree of fit is better.

Table 10 indicates that the regression equation is meaningful. The F value is an indicator of the significance of multiple regression. The results are all greater than 2.38. The original assumption is rejected; that is, the independent variables of the model are considered to have a sufficiently significant effect on the dependent variables.

The VIF in Table 11 is an indicator for common linear diagnosis. When $0 < \text{VIF} < 10$, the regression model is not collinear; $10 \leq \text{VIF} \leq 100$ is considered to indicate multi collinearity in the model. The VIF in the above table is less than 10; thus, the model does not exhibit multiple collinearity. Let the morbidity variable be X_1, the appropriate travel degree be X_2, the population mobility be X_3, and the promotion factor be X_4. The equation for Table 11 is sales = 2.346 X_1 + 1.094 X_2 − 1.122 X_3 + 2.295 X_4 − 4.192. Through the BPNN multivariate regression analysis model,

Table 5: Numerical table of the air pollution indexes.

Air pollution index	Evaluation criteria
[0, 50]	5
(50, 100]	4
(100, 200]	3
(200, 300]	2
(300, +∞]	1

Table 6: Numerical table of population mobility.

Population mobility	Evaluation criteria
Spring Festival, National Day	3
New Year's Day, Ching Ming, Labour Day, Dragon Boat Festival, Mid-Autumn Festival	2
Rest of the time	1

Table 7: Correlation analysis.

		Incidence of the disease	Suitable for going out	Population mobility	Promotion	Daily sales
Incidence of the disease	Pearson correlation	1	0.091	0.066	0.135	0.808
	Sig.(one-sided)		0.007	0.038	0	0
Suitable for going out	Pearson correlation	0.091	1	−0.129	−0.019	0.322
	Sig.(one-sided)	0.007		0	0.306	0
Population	Pearson correlation	0.066	−0.129	1	−0.04	−0.174
	Sig.(one-sided)	0.038	0		0.139	0
Promotion	Pearson correlation	0.135	−0.019	−0.04	1	0.238
	Sig.(one-sided)	0	0.306	0.139		0
Daily sales	Pearson correlation	0.808	0.322	−0.174	0.238	1
	Sig.(one-sided)	0	0	0	0	

Table 8: Importance of independent variables.

	Importance	Importance of standardization (%)
Incidence of the disease	0.382	100.00
Suitable for going out	0.220	57.70
Population mobility	0.175	45.70
Promotion	0.223	58.40

Table 9: Model summary.

	R	R^2	Adjustment R^2	Standard estimation error
Model	0.877	0.769	0.768	1.217

Table 10: Anova[a].

		Square sum	df	Mean square	F	Sig.
Model	Regression	3576.476	4	894.119	604.004	0.000e
	Residual	1074.712	726	1.48		
	Total	4651.187	730			

$K_1 = 0.7$ and $K_2 = 0.29$ are obtained. The expression of the combination model is $Q = 0.71\ Q_1 + 0.29\ Q_2$.

Then, the error rates of the neural network model, multivariate regression model, and combination model are obtained. Table 12 (the monthly summary sales) shows that the error rate of the combined model is generally less than that of each model; that is, the accuracy is higher than that of the other two models, with the overall error level being less than 10%.

After examining the accuracy and stability of the model, using formulas (3) and (4), we can obtain the replenishment quantity of the cold medicine.

4. Results and Discussion

4.1. Drug Replenishment Process. Refer to [23], the pharmacy replenishment process is shown in Figure 3. The wholesale association that manages pharmacies pays

Table 11: Result test chart.

Model	Non-standardized coefficient		Standard coefficient	t	Sig.	Correlation			Collinear statistics	
	B	Standard error	Applicable edit			Zero order	Partial	Part	Toler-ance	VIF
(N)	−4.192	0.485		−8.641	0					
Incidence of the disease	2.346	0.054	0.782	43.075	0	0.808	0.848	0.768	0.966	1.035
Suitable for going out	1.094	0.087	0.229	12.641	0	0.322	0.425	0.226	0.972	1.029
Population mobilety	−1.122	0.106	−0.191	−10.558	0	−0.174	−0.365	−0.188	0.974	1.026
Promotion	2.295	0.321	0.129	7.15	0	0.238	0.256	0.128	0.978	1.023

Table 12: Error rate of the model.

Month-year	Actual value	BPNN	Regression	Combine	Error rate (BPNN) (%)	Error rate (regression) (%)	Error rate (combined) (%)
Jan-16	251	236	247	239	−6	−2	−5
Feb-16	235	204	209	205	−13	−11	−13
Mar-16	217	228	228	228	5	5	5
Apr-16	157	140	149	143	−11	−5	−9
May-16	99	91	105	95	−8	6	−4
Jun-16	80	87	99	90	9	24	13
Jul-16	83	93	104	96	12	25	16
Aug-16	90	93	116	100	3	29	11
Sep-16	95	87	104	92	−8	9	−3
Oct-16	149	126	115	123	−15	−23	−18
Nov-16	199	170	168	169	−15	−16	−15
Dec-16	256	219	221	220	−14	−14	−14
Jan-17	245	210	218	212	−14	−11	−13
Feb-17	222	220	226	222	−1	2	0
Mar-17	195	204	205	204	5	5	5
Apr-17	140	149	159	152	6	14	9
May-17	90	89	102	93	−1	13	3
Jun-17	84	88	107	94	5	27	11
Jul-17	86	93	110	98	8	28	14
Aug-17	137	141	146	142	3	7	4
Sep-17	144	156	161	157	8	12	9
Oct-17	145	132	141	135	−9	−3	−7
Nov-17	194	167	173	169	−14	−11	−13
Dec-17	245	241	245	242	−2	0	−1

attention to the drug balance information of the pharmacies in real time. If the number of drugs is higher than the replenishment point, then replenishment is unnecessary. When the number of drugs is lower than the replenishment point, an alarm is automatically sent to the wholesaler for replenishment. Artificial intelligence technology exists inside the information system. The intelligent calculation of the replenishment amount is then carried out by the staff to pick up the goods according to the replenishment amount and arrange the goods to depart the warehouse. The drugs are then delivered to pharmacies that need to be replenished to complete the pharmacy replenishment process. Then, one goes to the next replenishment point and repeats the above steps, thus restarting the drug replenishment services cycle.

4.2. Model Simulation Results. In order to test the accuracy and error of the model, this article uses the 2018 data of a cold medicine in one pharmacy. At the beginning of the first day of 2018, the inventory of this cold medicine in the store was the same as the inventory limit, which was 130 boxes. Then, the simulation starts. Other known conditions are the same as those set in the previous chapter. By comparing the real data in 2018 with the replenishment data obtained by simulation, the results are shown in Figure 4 and Table 13.

Figure 4 reflects the comparison between the real data and the simulation data of the drug replenishment quantity in 2018. It can be seen from the figure that the drug replenishment quantity varies greatly in different seasons. Compare the real data and the model fitting data. It shows that the time with large errors is concentrated on the replenishment quantity in late January and early February, as well as in September and October. The forecast accuracy rate at other times is relatively high. From the error rate of the data in Table 13, the overall error is not very large. The absolute value of the error during each replenishment period does not exceed 20%,

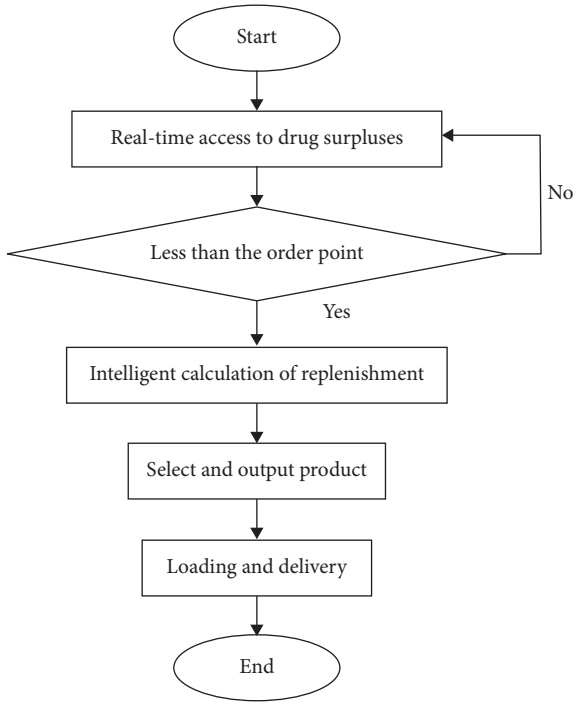


Figure 3: Drug replenishment process.

Table 13: Model simulation results.

Restock time	Error rate (real sales model result) (%)
Jan. (1)	5
Jan. (2)	−13
Feb. (1)	−17
Feb. (2)	−2
Mar. (1)	10
Mar. (2)	13
Apr. (1)	3
Apr. (2)	−3
May	13
Jun.	−3
Jul.	0
Aug.	8
Sep.	−13
Oct.	−16
Nov. (1)	−5
Nov. (2)	−3
Dec. (1)	−4
Dec. (2)	−3

and the average error rate is 7% and less than 10%. Obviously, the prediction accuracy is better. Better forecast months are June, July, August, November, and December. The month with the smallest error is July, and the premeasured and actual sales quantities are the same. The months with the largest errors between forecast and actual values are the second half of January, March, May, September, and October. At present, this article speculates that the main factors affecting the accuracy of sales forecasts are seasonal changes (alternating seasons) and the time of holidays because these factors are caused by sudden changes in drug sales (rapid decline or surge). The

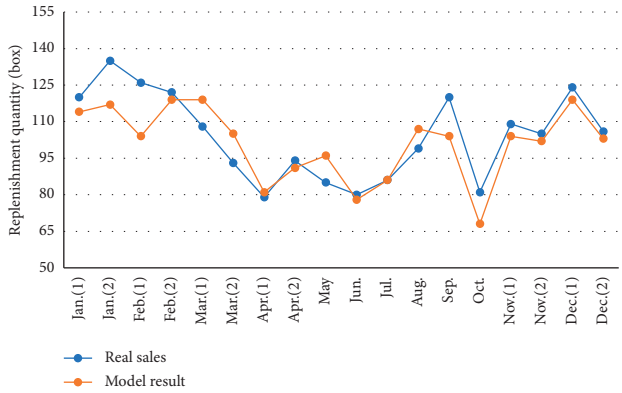


Figure 4: Model simulation comparison.

prediction model cannot quickly adapt to changes in actual values, and the prediction error is large.

5. Conclusions

A replenishment model is established to reasonably predict the replenishment quantity of drugs in a pharmacy, which solves the inventory shortage problem of the replenishment quantity prediction technology in this paper. For example, human experience alone determines the amount of drug replenishment; otherwise, the same amount is used each time. Specifically, first, the empirical formula is used to establish the replenishment model, and the safety stock and replenishment point can be obtained from the known historical sales data and other known data related to the replenishment. With the historical inventory turnover rate, the number of days sold for each replenishment, such as the next week, two weeks, or a month, is determined. That number is then used to forecast the demand over the coming days when the drugstore inventory is at or below the replenishment point. Then, the replenishment quantity is forecast by collecting and processing the known influencing factors and historical daily sales data related to the sales quantity of the drug, and the BPNN and regression analysis are used to carry out the regression analysis. Subsequently, the future demand is forecasted by the weighted combination model, and the exact replenishment result is obtained. Finally, the replenishment quantity is obtained by comparing with the known inventory upper limit.

BPNN multivariate regression analysis is more reliable than a single model and has better predictions of the timing of seasonal changes or sharp changes in sales over holidays. This point in the last section for the model forecast accuracy test also confirmed the accuracy and validity of its forecast results. The average error rate of approximately 7% means that the proposed model can better forecast the future replenishment quantity. It is more suitable for the prediction of the drug replenishment quantity of a pharmacy, making the replenishment time more reliable, avoiding the phenomenon of stock shortage or backlogs, and improving the capital turnover of the pharmacy, thus enabling the

pharmacy to use more funds on other hot-selling drugs to maximize pharmacy profits.

However, this model also has its limitations; it is intuitively that BPNN has better prediction accuracy than the multiple regression model, but the prediction accuracy of BPNN is not as good as that of the multiple regression model when the seasons change. Although the multiple regression model has the above advantages, the prediction fitting degree of it is worse than that of BPNN when the drug demand is stable. Therefore, when the two types of time prediction models are combined, the prediction stability is better than that of BPNN but the prediction accuracy of each month may be not the best, and the specific limitation is mainly due to the following reasons. On the one hand, when the sales quantity regression analysis was conducted in the first phase, it was obtained by analysing the relationship between the factors that affecting sales quantity and sales. In some instances, some of the impact factors are not taken into account, leading to inaccurate results. Another aspect is that the factors that affect sales in the future are the results obtained through the predictions, some of which are based on history and some on technical analyses conducted by professional organizations. Errors are likely, such as in the prediction of future weather due to the long duration considered. Inaccurate phenomena may occur, and certain errors in the conclusions will be reached. If an emergency occurs, it may also affect the accuracy of the results. For example, due to a certain virus in a certain year, the number of people infected with the disease will increase, and the amount of replenishment will not be adjusted in advance by predicting the incidence rate, which will lead to the prediction. The result is far lower than the actual need for replenishment, which may cause a loss in profits. To prevent this from happening, this paper proposed, in the visual interface concept of the model, that the amount of drug resupply be adjusted by manual floating up and down by no more than 20 boxes to make the premeasurement of resupply closer to the actual sales quantity to prevent this kind of problem from occurring.

In addition, this paper has established a model for the individual drug category in the field of pharmacies. However, in practice, many kinds of drugs exist. Many times, they still need to be combined with other drugs. This situation needs to be further studied and resolved.

Data Availability

The data of drug sales and promotion time used in the experiment are from WangJing store of Beijing Jinxiang pharmacy, which is located in Nanhu Dongyuan, Beijing. In the experiment, a cold medicine of a certain brand was used for prediction. The other data source is as follows: the appropriate travel data are from the "2345 weather forecast" website (http://tianqi.2345.com/); incidence rate data are from the website of Beijing Center for Diseases Prevention and Control (https://www.bjcdc.org/); population mobility data are from "China government network" (http://www.gov.cn/).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

All authors contributed to the study conception and design. Topic guidance was performed by Hao Zhang. Material preparation, data collection, and analysis were performed by Jia-Hui Mu and Hao Zhang. The first draft of the manuscript was written by Jia-Hui Mu, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The study was supported by the Beijing Philosophy and Social Science Foundation Project (grant no. 17GLB013), Talent Foundation Project of Organization Department of Beijing Municipal Committee of the CPC in 2018 (grant no. 2018000026833ZS09), Science and Technology Innovation Service Capacity Provincial - Ministerial Scientific Research Platform Construction Social Science Provincial - Ministerial Scientific Research Platform Construction Project (grant no. 19008020111), Construction Project of Humanities and Social Sciences Research Center at Provincial and Ministerial Level (grant no. 19002020217), and National Natural Science Foundation of China (grant no. 91646120).

References

[1] J. R. Trapero, M. Cardós, and N. Kourentzes, "Empirical safety stock estimation based on kernel and GARCH models," *Omega*, vol. 84, pp. 199–211, 2019.

[2] R. Fildes, "Research issues in business forecasting," *Management Research News*, vol. 11, no. 4-5, pp. 2–5, 2013.

[3] C. C. Holt, "Author's retrospective on "Forecasting seasonals and trends by exponentially weighted moving averages"," *International Journal of Forecasting*, vol. 20, no. 1, pp. 11–13, 2004.

[4] A.-L. Beutel and S. Minner, "Safety stock planning under causal demand forecasting," *International Journal of Production Economics*, vol. 140, no. 2, pp. 637–645, 2012.

[5] H. Lee, S. G. Kim, and H.-W. Park, "Pre-launch new product demand forecasting using the Bass model: a statistical and machine learning-based approach," *Technological Forecasting and Social Change*, vol. 86, pp. 49–64, 2014.

[6] D. Mmereki, B. Li, M. U. Hossain, and L. Meng, "Prediction of e-waste generation based on grey model (1, 1) and management in Botswana," *Environmental Engineering and Management Journal*, vol. 17, no. 11, pp. 2537–2548, 2018.

[7] J. H. Pang, H. Zhao, F. F. Qin, X. B. Xue, and K. Y. Yuan, "A new approach for product quality prediction of complex equipment by grey system theory: a case study of cutting tools for CNC machine tool," *Advances in Production Engineering & Management*, vol. 14, no. 4, pp. 461–471, 2019.

[8] A. Antonis, "A wavelet smoothing method to improve conditional sales forecasting," *Journal of the Operational Research Society*, vol. 66, no. 5, pp. 832–844, 2015.

[9] P. Ramos, N. Santos, and R. Rebelo, "Performance of state space and ARIMA models for consumer retail sales

forecasting," *Robotics and Computer-Integrated Manufacturing*, vol. 34, pp. 151–163, 2015.

[10] N. Kriegeskorte and T. Golan, "Neural network models and deep learning," *Current Biology*, vol. 29, no. 7, pp. 231–236, 2019.

[11] A. Bahrammirzaee, "A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems," *Neural Computing and Applications*, vol. 19, no. 8, pp. 1165–1195, 2010.

[12] C. Karunanayake, M. B. Gunathilake, and U. Rathnayake, "Inflow forecast of Iranamadu reservoir, Sri Lanka, under projected climate scenarios using artificial neural networks," *Applied Computational Intelligence and Soft Computing*, vol. 2020, Article ID 8821627, 11 pages, 2020.

[13] G. Aldabbagh, D. M. Alghazzawi, S. H. Hasan et al., "Optimal learning behavior prediction system based on cognitive style using adaptive optimization-based neural network," *Complexity*, vol. 2020, Article ID 6097167, 13 pages, 2020.

[14] L. Wang, Z. Wang, H. Qu, and S. Liu, "Optimal forecast combination based on neural networks for time series forecasting," *Applied Soft Computing*, vol. 66, pp. 1–17, 2018.

[15] T.-M. Choi, Y. Yu, and K.-F. Au, "A hybrid SARIMA wavelet transform method for sales forecasting," *Decision Support Systems*, vol. 51, no. 1, pp. 130–140, 2011.

[16] C. Christodoulos, C. Michalakelis, and D. Varoutas, "Forecasting with limited data: combining ARIMA and diffusion models," *Technological Forecasting and Social Change*, vol. 77, no. 4, pp. 558–565, 2010.

[17] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 465–474.

[18] L. Wang, B. Wu, Q. Zhu, and Y.-R. Zeng, "Forecasting monthly tourism demand using enhanced backpropagation neural network," *Neural Processing Letters*, vol. 52, no. 3, pp. 2607–2636, 2020.

[19] R. Cui, D. Zhang, and A. Bassamboo, *Learning from Inventory Availability Information: Field Evidence from Amazon*, Social Science Electronic Publishing, New York, NY, USA, 2016.

[20] J. Foster, C. Deck, and A. Farmer, "Behavioral demand effects when buyers anticipate inventory shortages," *European Journal of Operational Research*, vol. 276, no. 1, pp. 217–234, 2019.

[21] M. S. Puga, S. Minner, and J. S. Tancrez, "Two-stage supply chain design with safety stock placement decisions," *International Journal of Production Economics*, vol. 209, pp. 183–193, 2019.

[22] Z. L. Zhang, Y. F. Wang, and Y. Li, "Inventory control model based on multi-attribute material classification: an integrated grey-rough set and probabilistic neural network approach," *Advances in Production Engineering & Management*, vol. 14, no. 1, pp. 93–111, 2019.

[23] M. C. Tang and S. F. Liu, "Analysis of enterprise order performance process under supply chain network environment," *Productivity Research*, vol. 8, pp. 106-107, 2007.

Hindawi
Complexity
Volume 2021, Article ID 6620425, 11 pages
https://doi.org/10.1155/2021/6620425

Research Article

Application Research of Key Frames Extraction Technology Combined with Optimized Faster R-CNN Algorithm in Traffic Video Analysis

Zhi-guang Jiang[1] **and Xiao-tian Shi**[2]

[1]*Hebei University of Science and Technology, Shijiazhuang 050000, China*
[2]*Shi Jiazhuang University of Applied Technology, Shijiazhuang 050081, China*

Correspondence should be addressed to Xiao-tian Shi; shixt@sjzpt.edu.cn

Received 13 December 2020; Revised 7 January 2021; Accepted 13 January 2021; Published 2 February 2021

Academic Editor: Abd E.I.-Baset Hassanien

The intelligent transportation system under the big data environment is the development direction of the future transportation system. It effectively integrates advanced information technology, data communication transmission technology, electronic sensing technology, control technology, and computer technology and applies them to the entire ground transportation management system to establish a real-time, accurate, and efficient comprehensive transportation management system that works on a large scale and in all directions. Intelligent video analysis is an important part of smart transportation. In order to improve the accuracy and time efficiency of video retrieval schemes and recognition schemes, this article firstly proposes a segmentation and key frame extraction method for video behavior recognition, using a multi-time scale dual-stream network to extract video features, improving the efficiency and efficiency of video behavior detection. On this basis, an improved algorithm for vehicle detection based on Faster R-CNN is proposed, and the Faster R-CNN network feature extraction layer is improved by using the principle of residual network, and a hole convolution is added to the network to filter out the redundant features of high-resolution video images to improve the problem of vehicle missed detection in the original algorithm. The experimental results show that the key frame extraction technology combined with the optimized Faster R-CNN algorithm model greatly improves the accuracy of detection and reduces the leakage. The detection rate is satisfactory.

1. Introduction

Intelligent transportation is based on smart transportation. It makes full use of the Internet of things, cloud computing, Internet, artificial intelligence, automatic control, mobile Internet, and other technologies in the transportation field and collects traffic information through high tech, which is effective for traffic management, transportation, and the public. All aspects of transportation, such as travel, and the entire process of transportation construction management are controlled and supported, so that the transportation system has the ability to sense, interconnect, analyze, predict, and control in regions, cities, and even larger time and space, so as to fully guarantee traffic safety and provide the efficiency of transportation infrastructure, the improvement of the operation efficiency and management level of the transportation system, the service of smooth public travel, and sustainable economic development.

From the 1970s to the 1980s, intelligent transportation was proposed as a concept, but it was limited by computing power and communication means, and its development speed was slow. ITS research was in the preparatory stage, so the research at this stage mainly focused on the core of the ITS system, vehicle navigation system, and route planning guidance. Since the end of the last century, with the great development of data transmission speed, computing power, and positioning technology, the development speed of intelligent transportation has been greatly increased. Some developed countries such as the United States, Japan, and Europe have turned their research perspectives to verify

intelligence through the establishment of some large-scale projects, the transportation concept, and, based on this, comprehensive research and development of supporting basic technologies. The monitoring system is an important part of road traffic management. The video is captured by camera equipment installed on both sides or above the road to meet people's needs for real-time monitoring of traffic scenes. A large amount of road monitoring video equipment is built, and massive video data is accumulated in the traffic video system, which puts forward higher requirements for the storage capacity, transmission bandwidth, data analysis, and abnormal situation identification of the system. The continuous improvement of video processing capabilities and smart video recognition capabilities is the key technology to realize smart video. These data have the characteristics of huge capacity and large amount of information. Traditional traffic data processing methods, processing architectures, and smart video recognition algorithms have been gradually out of time, and they cannot meet the processing needs of intelligent transportation big data, especially smart video recognition. Instead, big data-related technologies are required to conduct in-depth mining and development of relevant data and adopt more advanced recognition methods to realize data sharing and integration, to achieve the purpose of intelligent services.

2. Research Status

In order to improve the accuracy of traffic smart video recognition, this paper proposes the use of key frame technology set combined with Faster R-CNN vehicle detection algorithm to judge vehicles. Most existing works use CNN as the feature extraction method of video. They divide video segments, extract one or more video frames from the segments as input to CNN, and then fuse the features of the segments. Early research took a single frame as the input of CNN. This approach makes insufficient use of time information. Some studies try to use CNN to extract video features directly. Tran et al. [1] trained a large-scale 3DCNN network. This type of approach increases the dimension of CNN at different levels, and the video sequence is directly used as the input of CNN. Limited by the size of the convolutional network, 3DCNN cannot handle videos of variable length and essentially still cannot avoid video segmentation recognition. Donahue et al. [2] proposed the LRCN structure, based on the use of CNN to extract the features of independent video frames, and introduced LSTM (a more effective RNN structure) to fuse the features extracted from each frame. The training time and storage cost of LSTM and 3DCNN are relatively large. In order to extract time information, Simonyan and Zisserman [3] proposed a dual-stream method. This method uses dense optical flow as an auxiliary input and uses two independent convolutional neural networks to extract the features of a single frame of original image and multiple frames of optical flow image and merge them at the final scoring level. The dual-stream method usually uses the network used by the image recognition task, and the calculation scale is equivalent to the image recognition task. The introduction of optical flow has significantly improved the accuracy of behavior recognition, and dual-stream networks have become the mainstream. Most of the subsequent researches improved deep neural networks on the basis of dual-stream networks or used various methods to fuse features based on dual-stream network features. The TSN (Temporal Segment Networks) proposed by Wang et al. [4] is a powerful improvement of dual-stream networks, and most of the existing methods use this as a measurement standard. This framework introduces a pretraining model of optical flow; at the same time, a training method of predividing video segments and using pooling fusion between segments is proposed. However, the TSN structure uses uniform segmentation, ignoring the difference in the amount of information between segments. Some people try to use the idea of key frames to improve the effect of behavior recognition. Hu and Zheng [5] used the optical flow difference method to extract the key frames in the video and achieved certain results in the KTH data set. In terms of deep learning algorithms, Girshick et al. [6] introduced the convolutional neural network to the target detection task for the first time, using the (Selective Search, SS) method to select candidate regions, and then using CNN to extract features and attach the classifier to the volume Perform detection on the product feature map and finally return to adjust the final position of the detection frame. Compared with the traditional algorithm, the average accuracy of this algorithm on the PASCAL VOC2012 test set (Mean Average Precision, MAP) is improved by 30%. In 2015, He et al. [7] proposed SPP net, which uses a spatial pyramid pooling layer to reduce the size limit of convolutional neural networks. Girshick [8] also proposed Fast-RCNN based on the idea of pyramid pooling in SPP net. This network uses a kind of ROI pooling to solve the problem that candidate boxes of different sizes cannot be input to the detection network with the same length and combine the candidate regions. If it is marked on the Feature map, only one feature extraction is required for the image, which greatly speeds up the operation of the network. Ren et al. [9] proposed the Faster R-CNN algorithm, which uses the RPN network to select candidate regions, which further reduces the running speed of the network and improves the detection accuracy.

3. Processing Architecture Based on Big Data

The continuous expansion of the construction of intelligent transportation video processing has resulted in massive heterogeneous data of different types and structures, such as system data, video data, and detection data, forming big data and traditional traffic data processing. The method and technical architecture can no longer meet the processing requirements of intelligent big data. Therefore, it is necessary to use big data-related technologies to conduct necessary mining and development of videos to achieve data sharing, processing, and integration to achieve the purpose of system processing requirements. This article refers to the sharing management method of massive data proposed by Wang et al. [10–12] and proposes related solutions in combination with virtual technology and distributed storage technology in cloud computing [13, 14].

3.1. Parallel Computing Model Design. In the traditional sense, cloud computing is divided into service modes, which can be divided into private cloud, public cloud, and hybrid cloud. According to the needs of the smart transportation video system, the article proposes a parallel computing mode, which is divided into two components: distributed file system (DFS) and distributed computing system (DCS); this computing mode has the following characteristics: (1) the client has the characteristics of flexible joining or evacuating; (2) since the application of each node is consistent, it can be based on the division of labor and different command files are configured for different tasks; (3) the deployment is simple, and the computing scale and storage scale can be controlled arbitrarily; (4) the model hides details of parallel computing, data distribution, load balancing, etc., and users can realize flexible computing according to actual needs and flexible processing; (5) the model has strong storage and computing capabilities and is fully adapted to data processing and data storage of smart video; and (6) the model can easily implement smart video algorithms such as convolutional neural networks [15].

The parallel computing model is shown in Figure 1.

3.2. Design of Distributed File Processing Model. The file distribution system is a network server component that makes it easier for users to query and manage data on the network. Distributed file system is a way to combine files distributed on different computers into a single name space and make it more convenient to establish a single, hierarchical multiple file server and server sharing work on the network, the traditional processing method. It is a "root node-node" model. Although the system architecture is greatly simplified, the core "root node" is responsible for the task of managing and accessing all "child nodes." The network pressure and computational pressure are huge. If a failure occurs, the entire system will be paralyzed, as shown in Figure 2.

In order to solve the problems of high root node pressure and high system risk in traditional distributed file systems, a new model is proposed. This model is composed of data storage nodes and management servers. The data storage nodes are responsible for data storage and data management services. The management server is responsible for managing the service process, maintaining the currently registered data service process, and analyzing the data source address. The management server can be an arbitrary computer in the cloud, which is a dynamic distributed file system. The distributed computer system is suitable for a variety of data processing modes, including distributed computing workstations and computing clients, which execute task allocation processes and task execution processes, respectively, among which multiple task execution processes can run, and these task allocation and execution processes can all be deployed on any computer in the cloud, and the data service process is deployed on the machine that stores the data and is responsible for the distribution and reception of the machine's data. This model belongs to the Master-Worker two-tier structure. The Master is responsible for task decomposition, task assignment, and client-related work. Once the Worker is started, it first registers with the task assignment, and the process assigns tasks to perform corresponding functions. The data service process requests data or uploads data and reports to the task allocation process according to its current state, as shown in Figure 3.

3.3. Video Processing Architecture Design. The video processing architecture consists of three parts: (1) data service process processing (DataServer), (2) task allocation process processing (WorkStation), and (3) task execution process (WorkClient). Among them, the task execution process can run multiple times at the same time according to the actual situation. The executable programs of the three processing processes are the same, but the composition of the respective command execution files (CmdFile) is different. The data processing task allocation process uses dynamic scheduling to allocate tasks. First, assign tasks to them according to the currently registered task execution process. Once the task is completed, it will apply for a new task to the task allocation process. At the same time, the assignable task execution process can be changing at any time, and the task execution process can also be added and withdrawn at any time. Since the execution process communicates with the data server separately, the communication overhead can be greatly reduced, and the additional overhead caused by management allocation can also be reduced. The data service process is deployed in the video data storage server or storage server array, communicates with the task execution process, distributes corresponding data according to the needs of the task execution process, and receives and stores the data processed by the task execution process. The task of the task allocation process is to analyze the task of video data processing and the allocation task. After the task is analyzed, it can also be used as a task execution process. The task execution process is mainly to perform a certain video processing subtask assigned by the task allocation process, as shown in Figure 4.

In data storage, in order to maintain storage flexibility, video data can be stored separately or stored in a server, and any computer in the cloud can become the management terminal. The relevant calculation process is as follows: (1) run the management service process; (2) start the data service process of the site machine where the data involved in the calculation is located; (3) register the data service process with the management service process and submit the managed data source to the local machine; (4) the data client submits the required data network data to the management service process, and the management service process parses the relevant parameters to the data client; and (5) the data client requests data processing services from the site machine based on these parameters.

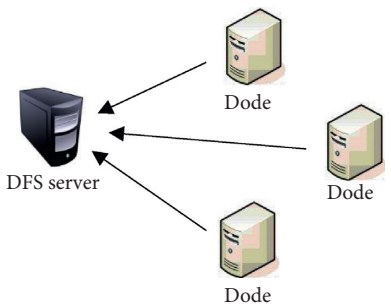


Figure 1: Parallel computing model based on cloud computing.

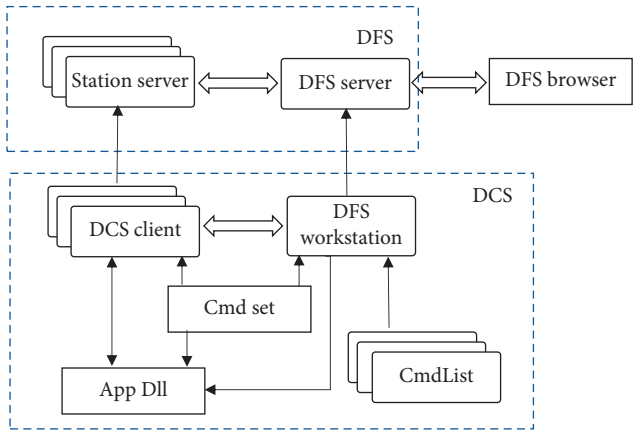


Figure 2: Structure of a distributed file system.

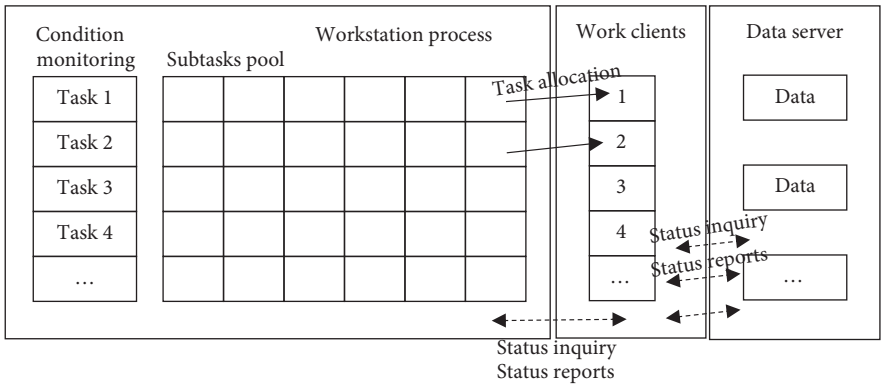


Figure 3: Distributed computing system architecture.

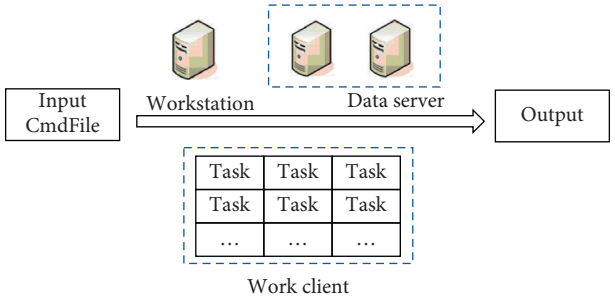


Figure 4: Video processing architecture.

4. Application of Video Key Frames Extraction Technology Based on Deep Learning

At present, in the main smart video surveillance field, pattern recognition technology is usually used. Common algorithms and detection methods include face recognition, anomaly recognition, and motion recognition, but most of these solutions are designed for specific application scenarios. The general scene recognition is not satisfied enough. Therefore, this research proposes a video key frame extraction and retrieval scheme based on deep learning. The key frame of the video represents the most significant feature of each shot of the video. Therefore, accurately extracting the key frame of each shot can effectively reduce the processing time of retrieval and improve the accuracy of retrieval. This scheme is divided into two steps. The first is to design an adaptive key frame recognition scheme and extraction scheme, and the second is to design a key frame retrieval algorithm based on convolutional neural network.

4.1. Video Layered Structure and Key Frames. The video can be divided into several scenes, each scene is divided into several shots, and each shot includes several key frames. Its structure is shown in Figure 5.

4.2. Shot Fragment Detection Algorithm. Video contains information such as spatial domain, time domain, plot, and features. Directly extracting and indexing features of the video are extremely complex work and consume a lot of storage space and computing time. If the accuracy of the extracted key frames is low, it will directly have an adverse effect on content-based video retrieval, video recognition, and scene analysis. As mentioned earlier, many researchers have proposed many key frame extraction schemes, but these schemes still have shortcomings. For example, there are many algorithmic solutions that select the first frame of each shot as the key frame, but this solution is easy to lose a large amount of visual information of the lens, and the randomness is strong, and the scale is not easy to grasp; there are some solutions to select key frames by enumerating and comparing each frame of the lens. The pressure on computing power and storage capacity is also great. According to the key frame extraction technology that needs to have the characteristics of high accuracy and fast calculation speed, combined with the scheme proposed by Liang and Wen [16], this paper proposes a new algorithm, as follows:

Input: video sequence, output: lens S_1, S_2, S_3, ..., S_n.

4.3. The Key Frame Extraction Algorithm of the Lens. Each shot contains many repeated frames, so there is no need to process each frame in the shot. First, extract the summary information of each shot of the video, and the extracted key frames should contain the most salient features. The key frame extraction algorithm of the shot is as follows:

The input is video footage, and the output is key frame collection.

FOREACH Frame f_n

Choose $f_1 f_2$ /*Choose $f_1 f_2$*/

FOREACH Blockb_n /*b_n is non-overlapping block of 16×16*/

Choose $b_1 b_2$

Wavelet exchange of $b_1 b_2 : F_{ij}(x, y) = X(x, y) f_{ij}(x, y) X^{-1}(x, y)$

Calculate the distance between wavelet transform blocks:

$$L2_{ij} = \sqrt{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (|F_{ij}(x, y) - F_{ij-1}(x, y)|)^2}$$

Save the distance to the vector

Calculate the average distance of the distance vector

IF $L2_{ij} \le LT$, /*LT is lens threshold*/

Conditional frames belong to the same lens

FOREACH lens st_n

FOREACH Frame f_n

Calculate the average of each frame and save it in a vector

Count the minimum, maximum, and mean values of the mean vector

Select the frame closest to the average value as the key frame

Firstly, calculate the average value of all frames in the shot, and use the frame with the average value closest to the vector average as the key frame, thereby realizing adaptive key frame extraction.

5. Implementation of Vehicle Detection in Traffic Monitoring Video Based on Deep Learning

5.1. Related Theoretical Basis. The overall architecture of the Faster-RCNN network is shown in Figure 6. The main functions of each layer are as follows:

(1) Conv layers extract feature maps: as a CNN network target detection method, Faster R-CNN first uses a set of basic conv + relu + pooling layers to extract the feature maps of the input image, which will be used in the subsequent RPN layer and fully connected layer.

(2) RPN (Region Proposal Networks): the RPN network is mainly used to generate region proposals. First, a bunch of anchor boxes are generated. After cutting and filtering them, Softmax is used to determine whether the anchors belong to the foreground or the background, that is, object or not object, so this is a two-category, at the same time, another branch bounding box regression modifies the anchor box to form a more accurate proposal (note: the more accurate here is relative to the next box regression of the fully connected layer).

(3) ROI pooling: this layer uses the proposals generated by RPN and the feature map obtained from the last layer of VGG16 to obtain a fixed-size proposal feature map. After entering it, it can use the full

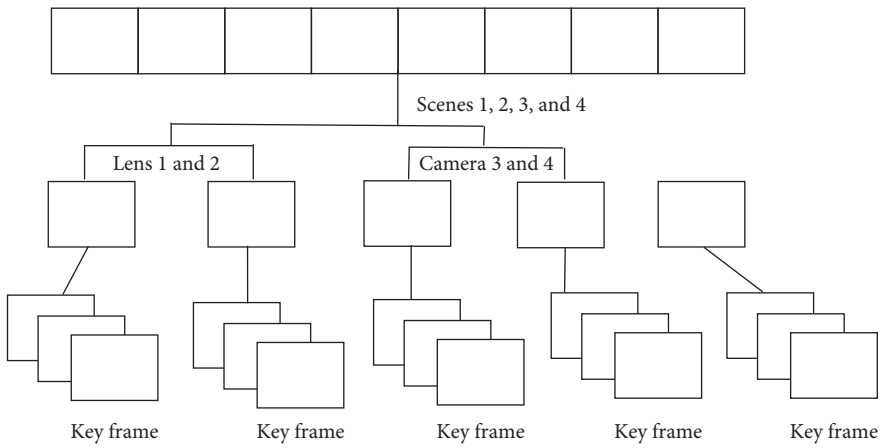


Figure 5: Video layered structure diagram.

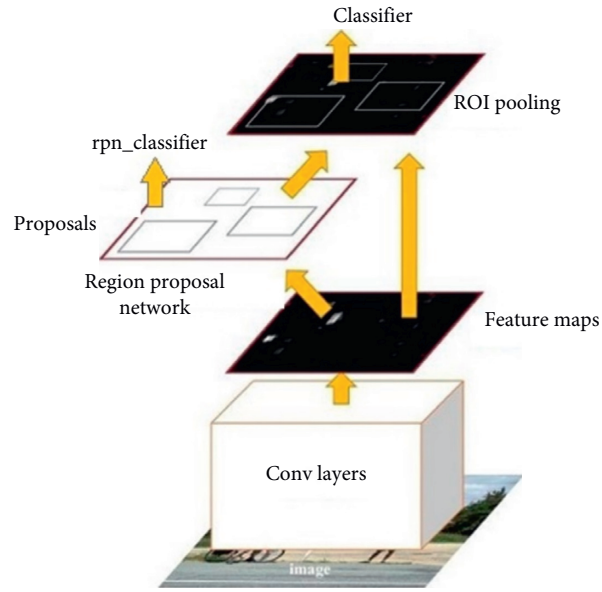


Figure 6: The overall architecture of the Faster R-CNN network.

connection operation to perform target recognition and positioning.

(4) Classifier: the ROI pooling layer will be formed into a fixed-size feature map for full connection operation, Softmax is used to classify specific categories, and at the same time, L1 Loss is used to complete the bounding box regression operation to obtain the precise position of the object.

When Faster R-CNN is applied to a traffic video surveillance system, high-resolution images may cause redundant feature information.

5.2. Vehicle Inspection Model Design. This research uses deep learning methods to solve the problem of vehicle detection under surveillance video [17]. The vehicle detection process is divided into two stages: training and detection. The execution steps of the training phase are as follows: (1) extract the training sample set, and make the sample training set according to the PASCAL VOC data set format and (2) input the sample training set into the neural network for training, and after multiple iterations, the trained vehicle detection network is obtained; the execution steps in the detection stage are as follows: (1) input the image directly into the trained neural network, and obtain the specific position of the outer frame of the vehicle and mark it on the original image and (2) output the vehicle detection result, as shown in Figure 7.

In traffic surveillance videos, high-resolution images are likely to cause feature information redundancy, and the use of nonmaximum values to suppress NMS when vehicles overlap most of the time can easily cause the detection frame to be lost. In order to solve this problem, we can consider using residual network technology to optimize the feature extraction layer, using hole convolution to filter redundant features, and using Soft-NMS to filter candidate frames. The structure is shown in Figure 8. Based on this model, it is divided into four components: feature extraction layer,

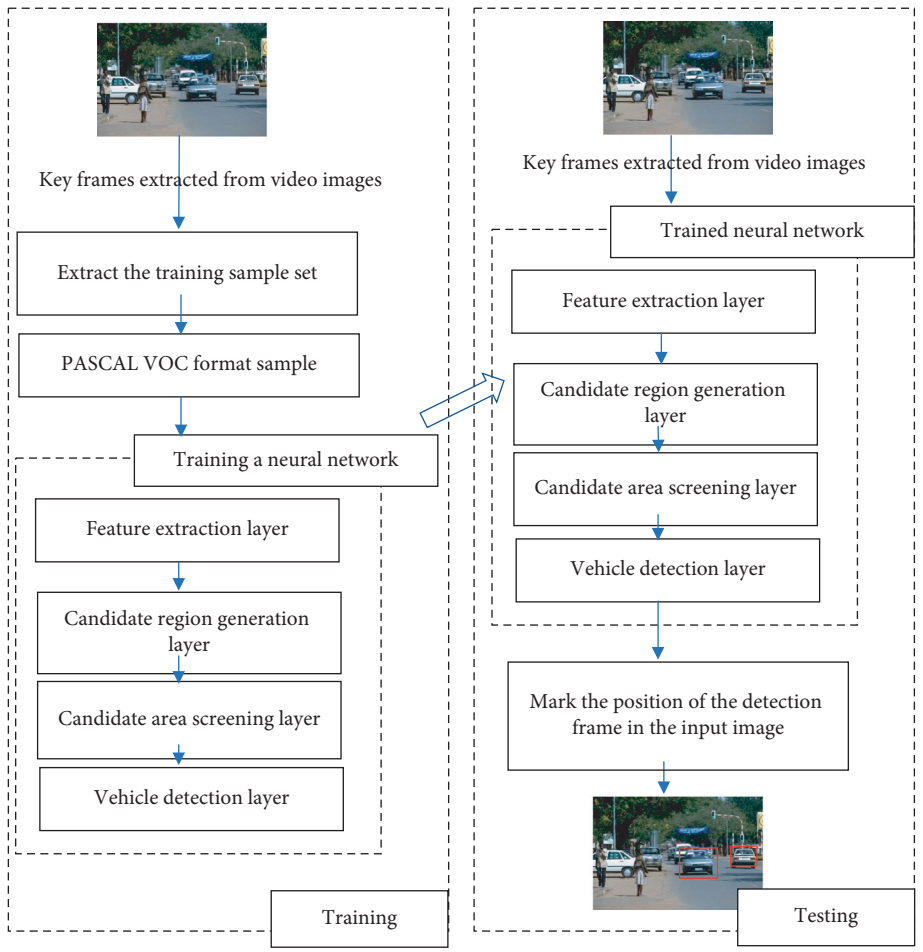


Figure 7: Vehicle detection algorithm flow based on faster R-CNN.

candidate region generation layer, candidate region screening layer, and vehicle detection layer.

(1) Feature extraction layer: perform feature extraction on the input image to generate a feature map

(2) Candidate region generation layer: use 3×3 hole convolution with expansion coefficient $r = 2$ to filter redundant features, and add anchor mechanism to generate initial candidate regions;

(3) Candidate region screening layer: compared with NMS, Soft-NMS is a softer screening criterion for candidate frames. Therefore, Soft Nonmaximum Suppression (Soft-NMS) is used for coarse screening of initial candidate regions.

(4) Vehicle detection layer: perform ROI pooling on the candidate area, realize a fixed length and output to the fully connected layer, and then connect to Soft-NMS again to screen the vehicle detection frame, and finally output the vehicle detection result.

5.3. Loss Function Design. The function of the loss function is to adjust the original model according to different environments to ensure the accuracy of data processing. The neural network is trained by defining a multitask loss function, such as the following:

$$L(\{p_i\}\{l_i\}) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(l_i p l_i^*), \tag{1}$$

where $N_{\text{cls}}, N_{\text{reg}}$, and λ balance the normalized weights of classification loss and regression loss and I is the index of the i-th candidate frame in small batch processing.

The probability is that the i-th candidate box is the target. If the i-th candidate box is a candidate target, then $p_i^* = 1$; otherwise, $p_i^* = 0$. The classification loss function and regression loss function are defined as formulae (2) and (3):

$$L_{\text{cls}}(p_i p_i^*) = -\log[p_i p_i^* + (1 - p_i^*)(1 - p_i)], \tag{2}$$

$$L_{\text{reg}}(t_i t_i^*) = R(t_i - t_i^*), \tag{3}$$

where R is the smooth$_{\text{L1}}$ function. $t_i = \{t_x, t_y, t_w, t_h\}$ is a vector-prediction parameterized candidate frame

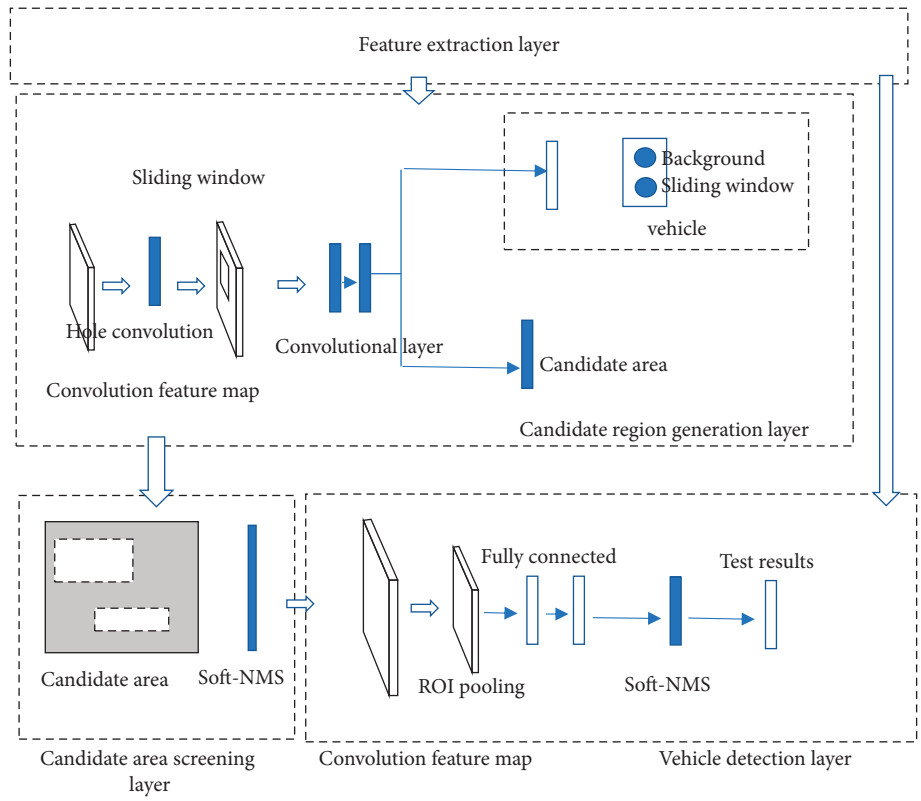


Figure 8: Structure diagram of neural network model.

coordinates and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}$ is the coordinate vector of real boundaries.

t_i and t_i^* are defined as follows:

$$
\begin{aligned}
t_x &= \frac{(x - x_j)}{w_j}, \\
t_y &= \frac{(y - xy_j)}{h_j}, \\
t_w &= \log\left(\frac{w}{w_j}\right), \\
t_h &= \log\left(\frac{h}{h_j}\right), \\
t_x^* &= \frac{(x^* - x_j)}{w_j}, \\
t_y^* &= \frac{(y^* - y_j^*)}{h_j}, \\
t_w^* &= \log\left(\frac{w^*}{w_j}\right), \\
t_h^* &= \log\left(\frac{h^*}{h_j}\right),
\end{aligned}
\tag{4}
$$

where (x, y), (x_j, y_j), and (x^*, y^*) are the forecasting area, candidate area, and official regional centre coordinates and (w, h), (w_j, h), and (w^*, h^*) are the width and height of predicted regions, candidate regions, and formal regions, respectively.

5.4. System Execution Process. The vehicle detection algorithm can be divided into two stages: training and detection. The main steps are as follows.

5.4.1. Training Part. The training process is shown in Figure 9.

5.4.2. Detection Section. The flow of the detection part is shown in Figure 10.

6. Implementation of Vehicle Detection in Traffic Surveillance Video Based on Deep Learning

Considering that there are different conditions in nature, such as daytime, night, rainy days, and traffic congestion, the research team conducted multiple sets of comparative experiments. The experimental results are shown in Table 1.

In the daytime environment, the accuracy of the system algorithm is more than 90%, and the effect is good. The accuracy of the system algorithm is more than 70% in traffic jams and rainy night environments, which is basically within the available range. From the comprehensive results, the algorithm designed in this paper is basically satisfactory.

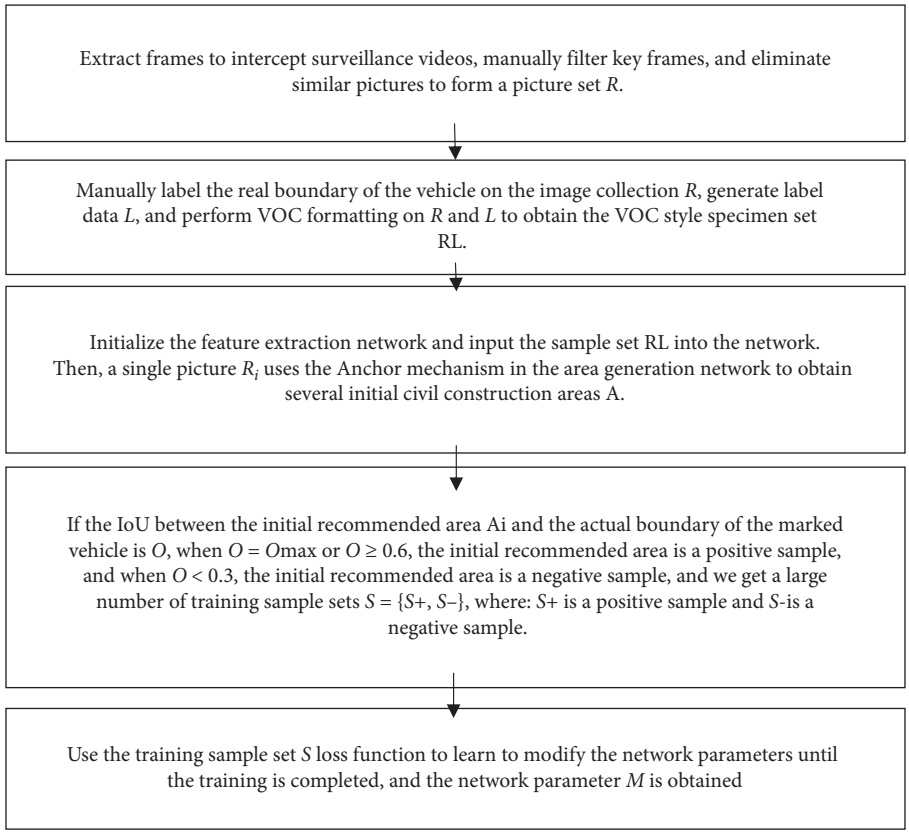


Figure 9: System implementation flow-training section.

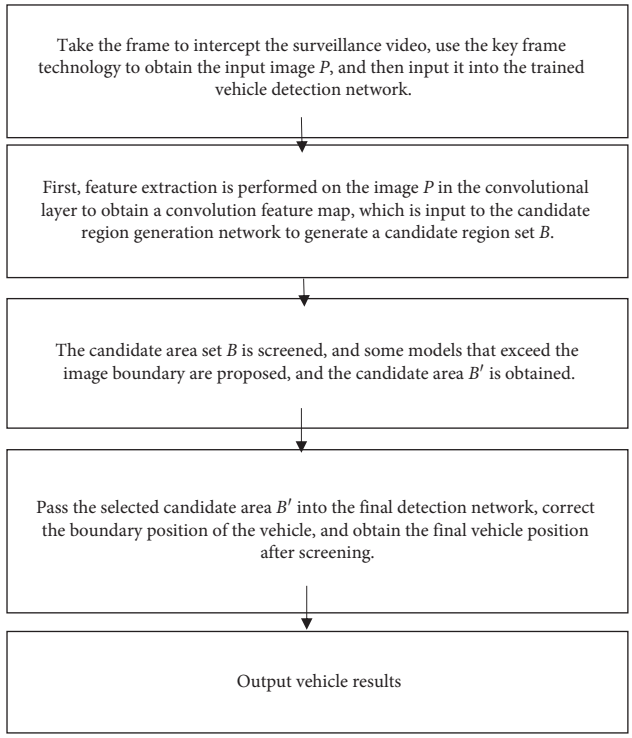


Figure 10: System implementation flow-detection section.

Table 1: Statistics of algorithm accuracy in different environments.

	Daytime	Daytime with rain	Cloudy day	Night with light	Night with rain	Traffic congestion	Synthesis
Accuracy (%)	91.2	85.3	88.6	80.6	70.6	71.6	79.9

Table 2: Comparison of accuracy between this algorithm and other algorithms.

	Accuracy (%)						
	Daytime	Daytime with rain	Cloudy day	Night with rain	Night with rain	Traffic congestion	Synthesis
This algorithm	91.2	85.3	88.6	80.6	70.6	71.6	79.9
R-CNN	77.5	71.6	76.3	73.2	63.5	62.1	69.6
SA-FRCNN	73.6	68.9	70.5	69.8	60.3	59.8	66.5
DPM	54.4	50.3	52.6	48.4	42.3	43.2	47.7

It can be seen from Table 2 that the performance of the traditional vehicle detection algorithm DPM is poor, especially under complex conditions, and the accuracy is less than 50%. R-CNN and SA-FRCNN are given to deep learning algorithms. The accuracy is much higher than that of traditional DPM, and the average accuracy is close to 70%. However, the accuracy of the algorithm in this study is significantly higher than that of the traditional algorithm.

7. Conclusion

Aiming at the characteristics of complex traffic video surveillance scenes and high resolution of single-frame video images, in order to improve the accuracy and time efficiency of the video retrieval scheme, a deep learning-based video key frame extraction and video retrieval scheme is proposed, which is combined with Faster R-CNN. The improved algorithm of R-CNN vehicle detection is used in intelligent traffic video analysis. First, add a hole convolution to the network to filter out the redundant features in the high-resolution video image, and then replace the original NMS mechanism with Soft-NMS to adapt to the overlap of vehicles, making it more suitable for traffic monitoring video vehicle detection. The experimental results show that the improved model improves the accuracy of detection and reduces the missed detection rate, and the experimental results are satisfactory.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful for the financial support provided by the Research and Practice Project of Higher Education Teaching Reform in Hebei Province, China (2019GJJG604).

References

[1] D. Tran, L. Bourdev, R. Fergus et al., "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the Computer Vision (ICCV), 2015 IEEE International Conference on (S1550-5499)*, pp. 4489–4497, IEEE, Santiago, Chile, December 2015.

[2] J. Donahue, L. A. Hendricks, M. Rohrbach et al., "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 1, no. 4, pp. 568–576, 2014.

[4] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," *Acm Transactions on Information Systems*, vol. 22, no. 1, pp. 20–36, 2016.

[5] Y. Hu and W. Zheng, "Human action recognition based on key frames," in *Advances in Computer Science and Education Applications (S1865-0929)*, pp. 535–542, Springer Berlin Heidelberg, Berlin, Germany, 2011.

[6] R. Girshick, J. Donahue, T. Darrell et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, OH, USA, June 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[8] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Las Condes, Chile, December 2015.

[9] S. Ren, K. He, R. Girshick et al., "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2015.

[10] C. L. Wang and Z. S. Zhang, "Design of large-scale traffic video processing frameworks based on private cloud," *Computer Engineering and Applications*, vol. 53, no. 21, pp. 254–257, 2017.

[11] M. Armbrust, A. Fox, R. Griffith et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[12] C. Chen, J. Lin, X. Wu, J. Wu, and H. Lian, "Massive geospatial data cloud storage and services based on NoSQL database technique," *Journal of Geo-Information Science*, vol. 15, no. 2, pp. 166–174, 2013.

[13] L. L. Qin, N. W. Yu, and D. H. Zhao, "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video," *Tehnicki Vjesnik*, vol. 25, no. 2, pp. 528–535, 2018.

[14] L. L. Qin and L. H. Kan, "Application of video scene semantic recognition technology in smart video," *Tehnicki Vjesnik*, vol. 25, no. 5, pp. 1429–1436, 2018.

[15] H. Qian and L. L. Qin, "The design of intelligent transportation video processing system in big data environment," *Special Section on Big Data Technology and Applications in Intelligent Transportation*, vol. 8, pp. 13769–13780, 2020.

[16] J. S. Liang and H. P. Wen, "Key frame abstraction and retrieval of videos based on deep learning," *Control Engineering of China*, vol. 26, no. 5, pp. 965–970, 2019.

[17] N. Bodla, B. Singh, R. Chellappa et al., "Soft-NMS—improving object detection with one line of code," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

Hindawi
Complexity
Volume 2021, Article ID 6643131, 17 pages
https://doi.org/10.1155/2021/6643131

Research Article

Solving a Joint Pricing and Inventory Control Problem for Perishables via Deep Reinforcement Learning

Rui Wang,[1] **Xianghua Gan**,[1] **Qing Li**,[2] **and Xiao Yan**[1]

[1]*School of Business Administration, The Southwestern University of Finance and Economics, Chengdu, Sichuan, China*
[2]*China Construction Bank, Hengshui Branch, Hengshui, Hebei, China*

Correspondence should be addressed to Xianghua Gan; ganx@swufe.edu.cn

Received 24 October 2020; Revised 6 January 2021; Accepted 12 January 2021; Published 30 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

We study a joint pricing and inventory control problem for perishables with positive lead time in a finite horizon periodic-review system. Unlike most studies considering a continuous density function of demand, in our paper the customer demand depends on the price of current period and arrives according to a homogeneous Poisson process. We consider both backlogging and lost-sales cases, and our goal is to find a simultaneously ordering and pricing policy to maximize the expected discounted profit over the planning horizon. When there is no fixed ordering cost involved, we design a deep reinforcement learning algorithm to obtain a near-optimal ordering policy and show that there are some monotonicity properties in the learned policy. We also show that our deep reinforcement learning algorithm achieves a better performance than tabular-based Q-learning algorithms. When a fixed ordering cost is involved, we show that our deep reinforcement learning algorithm is effective and efficient, under which the problem of "curse of dimension" is circumvented.

1. Introduction

The inventory control of perishables has received increasing attention from the business community and academia. According to a report released by the Food Market Institute (2012) in the United States, as of 2005, the total sales of perishables accounted for more than half of sales in supermarkets and grocery stores in the US, and this proportion is still increasing. Meanwhile, losses due to the deterioration of perishables also account for a large proportion of the total retail cost. Besides, pricing is also an important and effective lever for the retail industry to manage the profitability of perishables. As shown in Karaesmen et al. [1] and Chen et al. [2], a firm's profit increases significantly by dynamic adjustment of prices of perishables according to the availability of the inventory and the remaining lives of perishables.

In this research, we study a joint pricing and inventory control problem for perishables in a finite planning horizon. Demand in each period depends on the current price and satisfies a Possion distribution. The problem of inventory control for perishables is usually more difficult than the one for nonperishables, in which the inventory state can be represented by a single variable. The state of perishables has to be recorded by a vector to account for items with different lifetimes, which makes the analytical studies much more difficult. As a fixed ordering cost can make the problem even more difficult in a dynamic setting, few studies consider it due to the tractability in analysis.

One main contribution in this research is that we consider a fixed ordering cost in our model. We study both the backlogging case and the lost-sales case and allow for positive lead time. Our goal is to find a near-optimal ordering and pricing policy to maximize the expected profit in the planning horizon. This problem is hard to analyze by traditional dynamic programming approach in the inventory control literature. Therefore, we use a reinforcement learning approach to solve the problem.

In the literature, there have been a few papers that study inventory control problems with reinforcement learning, such as Charharsooghi et al. [3], Dogan et al. [4], and Kara et al. [5]. Unlike these papers which use Q-learning, we take a deep reinforcement learning approach and show that

it outperforms Q-learning models that do not use neural networks. The outperformance of deep reinforcement learning has also been shown by Ke et al. [6] and Shihab et al. [7] for complex problems.

In this paper, we set up deep reinforcement learning models to study the joint pricing and inventory control problem of perishables. We adopt a FIFO (first-in-first-out) policy in this study. When there is no fixed ordering cost involved, we show that the fixed pricing strategy is dominated by the dynamic pricing strategy, under which the price can be adjusted according to the availability of inventory and the lives of remaining items. We set up a benchmark based on realized demand for this no fixed ordering cost case and show that our designed deep reinforcement learning methods achieve a better performance than tabular-based Q-learning. We also find some monotonicity properties in our learned policies; our learned order quantity is nonincreasing in inventory position or on-hand inventory and price decision is most sensitive to the oldest on-hand inventory. Moreover, in order to show the expansibility of the proposed algorithm, we extend the distribution of the demand and take an *additive* form in Chen et al. [2] where the customer demand depends on the price of current period plus an additive random term; finally, we obtain a near-optimal performance by our proposed deep reinforcement learning models. When the fixed ordering cost is taken into account in the joint pricing and inventory control system, we set up a performance upper bound based on the realized demand in each period in order to assess the performance. Through our proposed methods, we find convergent policies and critical values under which orders should be placed.

2. Literature Review

We review two streams of literature which are closely related to our research: traditional inventory control management for perishables and inventory control management with reinforcement learning.

2.1. Traditional Inventory Control Management for Perishables. There is a considerable literature devoted to dynamic inventory control for nonperishable products; see, for example, Presman and Sethi [8], Caliskan-Demirag et al. [9], Alp et al. [10], Almaktoom et al. [11], Azghandi et al. [12], Li et al. [13], and Gan et al. [14]. The dynamic inventory control for perishable products has not been widely studied in the literature. This is not to say that the literature does not realize the importance of the study of perishable studies.

Indeed, there are a number of papers devoted to the study of inventory decisions for perishable products. Nahmias and Pierskalla [15] studied a dynamic inventory control with a fixed lifetime, zero lead time, and uncertainty demand for perishable products. Nahmias [16], Fries [17], and Nahmias [18] studied the same problem, with multiple periods of lifetime and zero lead time, and their research studies are all to satisfy the same assumption that only products that exceed the life cycle will be abandoned, which is known as the first-in-first-out policy (FIFO), and this policy is widely used in the research of perishable retailing. And they proved that the optimal order quantity under different inventory ages is decreasing. Prastacos [19] reviewed some important theories and practices in blood inventory management and proposed that this kind of application can be extended to other perishable product inventory control problem. Ferguson and Koenigsberg [20] considered a two-period joint pricing and inventory control problem with a random lifetime, emphasizing and discussing the impact of competition between new inventory and surplus inventory over the previous period on inventory and pricing decisions for the first time interval. Chen et al. [21] used Pontryagin's maximum principle method to investigate the optimal policies for the pricing and replenishment of fashion apparel with short product lifecycles. Heuristic algorithms are also increasingly being used to address the problem of dynamic pricing and inventory control for perishables. Li et al. [22] proposed a base-stock/list-price heuristic policy to solve the problem of dynamic pricing and inventory control for a perishable product, assuming that the demand is a function of price and zero lead time. Li and Lu [23] studied a joint optimization of the price and order quantity of a perishable product and proposed a Minimax Regret algorithm. Li et al. [24] discussed a new dynamic pricing and inventory control scenario for perishables. New and old products cannot be sold at the same time. The seller can decide whether to discard the remaining inventory in the previous period, even though the lifetime may not be over. And they proposed a fractional programming heuristic algorithm to obtain a stable structural policy.

Chen et al. [2] is closely related to our research. They considered positive lead time and used the concept of L-convexity/concavity to analyze the problem and proposed a heuristic algorithm to solve the problem. However, the traditional approach used in their research is not able to solve the problem with a fixed ordering cost. By using neural networks with hidden layers to approximate state-action values, our deep reinforcement learning approach exploits the advantages of deep learning [25] and reinforcement learning and is shown to be effective and efficient to find the solution.

2.2. Inventory Control with Reinforcement Learning. In the literature, there have been few papers that study inventory control problems with reinforcement learning.

Giannoccaro and Pontrandolfo [26] studied the coordination of inventory policies adopted by different supply chain factors which are a major issue in supply chain inventory management, and they used a reinforcement learning approach to manage inventory decisions at all stages of the supply chain in an integrated manner and aimed at optimizing the performance of the whole supply chain. Chaharsooghi et al. [3] proposed an inventory control system based on reinforcement learning methods, which included uncertain delivery times and uncertain customer requirements to determine the ordering policy for each order point in the supply chain. Chaharsooghi et al. [3] used Q-learning to solve supply chain ordering management and

applied to the beer game. Jiang and Sheng [27] proposed a case-based reinforncement learning algorithm (CRL) for dynamic inventory control in a multiagent supply-chain system. They studied a multiagent simulation of a simplified two-echelon supply chain and showed the effectiveness of the method they proposed. Sui et al. [28] considered a Vendor-Managed Inventory (VMI) system where the supplier makes decisions of inventory decisions of inventory management for the retailer, and the retailer is not responsible for placing orders. Through a methodology based on reinforcement learning and numerical study, they show their approach can outperform the newsvendor. Zarandi et al. [29] presented a flexible fuzzy reinforcement learning algorithm where the value function is approximated by a fuzzy rule-based system and considered the problem of a fuzzy agent (supplier), that is, how to determine the amount of orders for each retailers based on their utility for supplier when its supply capacity is limited. Finally, the effectiveness of their proposed algorithm is proved by a simulation. Dogan et al. [4] used the Q-learning method to study an ordering and pricing policy in a multiretailer environment. Rana and Oliveira [30] use reinforcement learning methods to develop dynamic pricing strategies for interdependent perishable products or service. Kara and Dogan [5] used Q-learning and Sarsa reinforcement learning algorithms to study a dynamic inventory control issues for perishable products, with positive lead time and fixed lifetime. Our research further uses deep reinforcement learning to study this dynamic inventory control of perishable products.

The aforementioned studies investigate the inventory problem for nonperishable and perishable products and use the nondeep reinforcement learning methods. Compared to their problems, our problem focuses on the inventory control of perishables, which makes the problem much more difficult. We use neural networks to avoid the curse of dimensionality and show that our deep reinforcement learning model outperforms the traditional reinforcement learning models without using neural networks.

3. Model

We consider a periodic-review single-product inventory system over a finite horizon of T periods. The whole process can be defined as a Markov Decision Process. The decision maker is called the *agent*, and the thing it interacts with is called the *environment*. At each period (step) of a sequence of discrete time periods, $t = 1, 2, \ldots, T$, the agent and the environment interact; the agent selects the action denoted by A^t , and the environment responds to A^t and presents a new situation to the agent. At the end of the period, the agent receives a numerical *reward* denoted by R^{t+1}, $R^{t+1} \in \mathbb{R}$, in part as a consequence of its action. Throughout this paper, we let superscript t denote the period. More specifically, by superscript t, we mean the beginning of the period; we denote the end of period t, which coincides with the beginning of the next period as $t + 1$.

Customer demand, denoted by D^t, at the beginning of period t, is represented by a Poisson distribution with the parameter as d^t or $d^t(A^t)$ if the agent's action A^t at the beginning of period t changes the demand distribution and $d(\cdot)$ is a function of selling price p, strictly decreasing the selling price p. Let the product's finite *lifetime* be denoted by l, *variable cost* by c^t, and *leadtime* by L^t $(0 \le L^t < l)$. Let the *age* of an item be 0 by the time it is shipped to the agent, and its *residual lifetime* be $l - i$ when its age is i. When an item's age is greater than l, it has to be disposed. The inventory state, also known as the state of the agent, at the beginning of period t can be represented by a $(l-1)$-dimensional vector:

$$X^t = \left(x_1^t, \ldots, x_i^t, \ldots, x_{l-1}^t\right), \tag{1}$$

where x_i^t represents the level of inventory position of the items at the age of i. In particular, $x_o^t \equiv \sum_L^{l-1} x_i^t$ (when $L = 0$, $x_o^t \equiv \sum_1^{l-1} x_i^t$) is the level of on-hand inventory, and $x^t \equiv \sum_1^{l-1} x_i^t$ is the level of inventory position of all ages.

3.1. An Action. Here, action space A^t refers to the order quantity q^t and price decision p^t. The selling pricing p^t is restricted to an interval $[\underline{p}, \overline{p}]$. Based on the selling price p, the parameter of Poisson demand $d \in [\underline{d}, \overline{d}]$, where $\underline{d} = d(\overline{p}), \overline{d} = d(\underline{p})$:

$$A^t = \left(q^t, p^t\right). \tag{2}$$

3.2. Update Rule. Update rule, denoted by $h(\cdot)$, describes the update of the environment state. In our research, the supply state remains unchanged in each period $(M^{t+1} = M^t)$. The demand state in each period depends on selling price p. Last, we need to define the update rules for the inventory state.

The update rules for the inventory state X^t can be divided into two cases according to the unmet demand handing principle. We first consider the backlogging case. If $L = 0$ and $L = 1$, then $x_i^{t+1} = x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+$ for $i = 1$ and $x_i^{t+1} = (x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+)^+$ for $i = 2, \ldots, l-1$; although $L = 0$ and $L = 1$ have the same state transition rule, the reward function is different; if $L > 1$, then $x_i^{t+1} = x_{i-1}^t$ for $i = 1, \ldots, L-1$, $x_i^{t+1} = x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+$ for $i = L$, and $x_i^{t+1} = (x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+)^+$ for $i = L+1, \ldots, l-1$.

For the lost-sales case, if $L = 0$ and $L = 1$, then $x_i^{t+1} = (x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+)^+$ for $i = 1, \ldots, l-1$; if $L > 1$, then $x_i^{t+1} = x_{i-1}^t$ for $i = 1, \ldots, L-1$ and $x_i^{t+1} = (x_{i-1}^t - (D^t - (x_i^t + \cdots + x_{l-1}^t))^+)^+$ for $i = L, \ldots, l-1$.

3.3. Reward Function. In our study, our goal is to maximize the accumulative expected profit in the planning horizon, so our reward function R^{t+1} can be represented by the following form. We first consider the backlogging case. If $L = 0$, then

$$R^{t+1} = \begin{cases} p^t * D^t - c * q^t - h * \left(x_o^t + x_0^t - D^t\right) - \nu * \left(x_{l-1}^t - D^t\right)^+, & \text{if } x_o^t + x_0^t \geq D^t, \\ p^t * D^t - c * q^t - u * \left(D^t - x_o^t - x_0^t\right), & \text{if } x_o^t + x_0^t < D^t. \end{cases} \tag{3}$$

If $L > 0$, then

$$R^{t+1} = \begin{cases} p^t * D^t - c * q^t - h * \left(x_o^t - D^t\right) - \nu * \left(x_{l-1}^t - D^t\right)^+, & \text{if } x_o^t \geq D^t, \\ p^t * D^t - c * q^t - u * \left(D^t - x_o^t\right), & \text{if } x_o^t < D^t. \end{cases} \tag{4}$$

For the lost-sales case, if $L = 0$, then

$$R^{t+1} = \begin{cases} p^t * D^t - c * q^t - h * \left(x_o^t + x_0^t - D^t\right) - \nu * \left(x_{l-1}^t - D^t\right)^+, & \text{if } x_o^t + x_0^t \geq D^t, \\ p^t * \left(x_o^t + x_0^t\right) - c * q^t - u * \left(D^t - x_o^t - x_0^t\right), & \text{if } x_o^t + x_0^t < D^t. \end{cases} \tag{5}$$

If $L > 0$, then

$$R^{t+1} = \begin{cases} p^t * D^t - c * q^t - h * \left(x_o^t - D^t\right) - \nu * \left(x_{l-1}^t - D^t\right)^+, & \text{if } x_o^t \geq D^t, \\ p^t * x_o^t - c * q^t - u * \left(D^t - x_o^t\right), & \text{if } x_o^t < D^t. \end{cases} \tag{6}$$

where inventory carried forward to the next period incurs a unit holding cost h, unmet demand incurs a unit penalty cost u, and ν is unit disposal cost. When fixed ordering cost K is considered, reward function above will subtract K if order quantity is not 0.

The sequence of events in period t is as follows:

(1) Based on the environment state E^t, $E^t \equiv (D^t, M^t, X^t)$, the agent selects an *action* A^t. Note that A^t is a vector, including ordering and pricing decisions. The order will be delivered at the beginning of period $t + L$; when $L = 0$ the order is delivered immediately.

(2) During period t, demand D^t arrives, which is discrete and stochastic depending on the selling price p^t, and is satisfied by the on-hand inventory as much as possible by the agent. Unsatisfied demand is either backlogged or lost; the remaining inventory with positive lifetime can be carried over to the next period.

(3) At the end of period t, the agent receives a reward R^{t+1}, which depends on the environment state and action A^t.

(4) At the beginning of period $t + 1$, the agent receives an order (if any), and the environment state is updated to E^{t+1} according to the *update rule*.

$$h: \left(E^t, A^t\right) \longrightarrow E^{t+1}. \tag{7}$$

For this joint pricing inventory problem, we introduce the notations in Table 1.

In this paper, we assume as in Chen et al. [2] that $c \leq u/1 - \gamma$, which eliminates the incentive to intentionally carry the back orders. We also assume that items with different lifetimes are charged the same price and that the back orders are met at cost c at the end of each planning period.

4. Deep Reinforcement Learning Methods

The objective of reinforcement learning is to learn a policy π that achieves near-optimal accumulated reward for the agent. Q-learning [31] is one widely used value iterative reinforcement learning method where the expected total discount rewards of state-action pairs can be approximated by a Q-function table based on the bellman equation, as shown in Function 7. Q-learning also has obvious limitation, that is, when there is a large state space, it is impractical and inefficient to record all the states and actions. Mnih et al. [32] extends Q-learning to Deep Q-network (DQN) which uses a neural network to approximate the Q-function table. DQN updates the parameters of the neural network by minimizing the difference between the predicted Q-values and the target Q- values, where the target Q-values are estimated by current

Table 1: Notations grouped by the elements of the RL problem.

p^t	Price charged by the agent for each item
q^t	Order quantity
D^t	Customer demand ($D^t(p^t)$ if depending on p^t)
l	Lifetime of the product, $0 < l < \infty$
L	Order lead time, $L < l$
K	Fixed order cost
c	Variable cost per item
u	Penalty cost per unmet item
h	Holding cost for per item left
ν	Cost for each item disposed
x_i^t	Inventory position of age i, $0 \le i \le l-1$
x^t	Inventory position of all ages
x_o^t	On-hand inventory
R^{t+1}	The reward function of the agent

reward and predicted Q-values from the next state. Meanwhile, to avoid training instability caused by correlation between training data, a replay memory pool is used:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[R + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]. \tag{8}$$

As mentioned before, in our joint pricing and inventory control problem, the state of the agent is expressed by the inventory state X^t which integrates different ages and corresponding quantities. Here, the initial inventory state is X^0. In our proposed algorithm PAQ-DQN, there are two same neural networks with the same structure but different parameters θ and $\widehat{\theta}$, respectively. We adopt the fixed Q-targets' policy in standard DQN. The neural network that predicts Q-values has the latest parameters, while the neural network that predicts target Q-values uses the old parameters. Each neural network has two hidden layers, and there are 128 neurons in each layer, we use the ReLU as the activation function. In each time period, based on Q-values from the neural network, the ε-greedy policy will be executed to select an action from the action space which contains a combination of ordering and pricing. After receiving the reward from the environment, the target Q-values are estimated by current rewards and discounted predicted Q-values from the next state, as shown in equation (9). The parameters of the network θ are updated by minimizing the difference between the predicted Q-values and the target Q-values, as shown in equation (10). After a fixed number of steps, assign the value of parameter θ to $\widehat{\theta}$. The details of the algorithm named perishables integrate age and quantity deep Q-network (PAQ-DQN) are shown in Algorithm 1:

$$y^t = \begin{cases} R^{t+1}, & \text{if epoch terminates at step } t+1, \\ R^{t+1} + \gamma^* \max_{A'} \widehat{Q}\left(X^{t+1}, A'; \widehat{\theta}\right), & \text{otherwise,} \end{cases} \tag{9}$$

$$L(\theta) = E_{X_i^t, A_i^t, R_i^{t+1}, X_i^{t+1}}\left[\left(y^t - Q\left(X^t, A^t; \theta\right)\right)^2\right]. \tag{10}$$

The second reinforcement learning algorithm named perishables integrate age and quantity advantage actor-critic (PAQ-A2C). A2C is a method combining policy gradient and function approximation. Actor-critic (A2C) has two networks, one policy network, known as actor and used to output policy, and one value network, known as critic and used to evaluate the policy from actor. In our algorithm, both the policy network and value network have two hidden layers, and there are 128 neurons in each layer with the ReLU activation function. Especially, the activation function of the policy network output layer is the Softmax, which outputs the probability of each action being executed in the current state. In each time period, based on the current inventory state, an action will be executed by the policy network. After receiving the reward from the environment, the value network will evaluate this policy and output a *td_error*. The parameters of value network θ_v can be updated by Equation (11), where y^t is the target value calculated by equation (12). The policy network is updated by $\theta_p \leftarrow \theta_p + \alpha * \nabla_{\theta_p} J(\theta_p)$, where α is learning rate, and gradient $\nabla_{\theta_p} J(\theta_p)$ is shown in equation (13) where advantage function is estimated by equation (14). The details of the algorithm of proposed perishables integrate age and quantity advantage actor-critic (PAQ-A2C) are shown in Algorithm 2:

$$L(\theta_v) = \left[y^t - V^\pi\left(X^t; \theta_v\right)\right], \tag{11}$$

$$y^t = R^{t+1} + \gamma * V^\pi\left(X^{t+1}; \theta_v\right), \tag{12}$$

$$\nabla_{\theta_p} J\left(\theta_p\right) \approx \nabla_{\theta_p} \log \pi_{\theta_p}\left(A^t | X^t\right) \widehat{A}\left(X^t, A^t\right), \tag{13}$$

$$\widehat{A}\left(X^t, A^t\right) = R^{t+1} + \gamma * V^\pi\left(X^{t+1}; \theta_v\right) - V^\pi\left(X^t; \theta_v\right). \tag{14}$$

5. Experiments

In this section, we conduct simulation studies to evaluate the performance of our proposed reinforcement learning algorithms and investigate the positive effects of the proposed algorithms on the profit of dynamic pricing and the impacts of the key parameters. Ordering and pricing policy are also discussed in situation involving fixed ordering cost. In this experiment, we only show the discussions on the backlogging case, the discussions of the lost-sales case is carried out in Appendix.

The values of various parameters are set in Table 2. For simplicity, the value range of the price p and order quantity q are restricted to $[32, 37]$ and $[0, 31]$, respectively.

In the reinforcement learning method, the effect of hyperparameters on final performance is very important, so we need to set the variation rules for relevant parameters, exploration rate ε and learning rate α. We adopt ε-greedy policy here; ε is decreasing linearly, that is, search-then-convergence form in Darken et al. [33]:

$$\varepsilon_{\text{epoch}} = \frac{\varepsilon_0}{1 + y}, \tag{15}$$

where $y = \text{epoch}^2/\varepsilon_{\text{decay}}$, ε_0 is the initial value of the ε, and $\varepsilon_{\text{decay}}$ is the decay parameter.

(1) Initialize replay memory pool D to capacity N
(2) Use random weights θ to initialize the action-value function Q
(3) Initialize target action-value function $\widehat{Q}$ with weights $\widehat{\theta} = \theta$
(4) **For** $epoch = 1$ to number of $epochs$ **do**
(5) Reset the environment and initialize state X^0
(6) **for** $t = 1, T$ **do**
(7) With probability ε, select a random action A^t, otherwise select $A^t = \text{argmax}_{A^t} Q(X^t, A^t; \theta)$ (ϵ-greedy policy)
(8) Execute action A^t and observe reward R^{t+1} and X^{t+1}
(9) Store transition $(X^t, A^t, R^{t+1}, X^{t+1})$ in the replay memory pool D
(10) Set $X^{t+1} = X^t$
(11) Sample a minibatch of transitions $(X_i^t, A_i^t, R_i^{t+1}, X_i^{t+1})$, $\forall i = 1, \ldots, N$ from replay memory pool D
(12) Calculate the target Q-value by equation (9)
(13) Update the parameters of network θ by equation (10)
(14) Every C steps reset $\widehat{Q} = Q$
(15) **end for**
(16) **end for**

Algorithm 1: Perishables integrate age and quantity deep Q-network.

(1) Use random weights θ_p and θ_v to initialize the policy network and value network
(2) **for** $epoch = 1$ to number of $epochs$ **do**
(3) Reset the environment and initialize state X^0
(4) **For** $t = 1, T$ **do**
(5) Take action A^t based on action probability $\pi_{\theta_p}(\cdot|X^t)$
(6) Execute action A^t and observe reward R^{t+1} and X^{t+1}
(7) Update the parameters θ_v of the value network by minimizing the loss function equation (11)
(8) Estimate advantage function by equation (14)
(9) Update the policy network parameters $\theta_p \leftarrow \theta_p + \alpha_p \nabla_{\theta_p} J(\theta_p)$, where $\nabla_{\theta_p} J(\theta_p)$ is calculated by equation (13)
(10) Set $X^{t+1} = X^t$
(11) **end for**
(12) **end for**

Algorithm 2: Perishables integrate age and quantity advantage actor-critic.

Table 2: The parameter values.

Parameter	Values	Description
T	30	Periods of the planning horizon
l	{2,3,4}	Lifetime of perishables
x^l	0	Initial inventory position of all ages
γ	0.9	Discount factor
L	{0, 1, 2}	Lead time
K	{25, 50}	Fixed ordering cost
c	22.5	Unit variable cost
h	0.22	Unit holding cost
u	10.78	Unit penalty cost
v	10	unit disposal cost
d^t	$84 - 2p$	Function of selling price
ε_0	1	Initial exploration rate
α	{0.01, 0.001, 0.0001}	Learning rate
$\varepsilon_{\text{decay}}$	$\{1 \times 10^3, 1 \times 10^4, 1 \times 10^5\}$	Exploration rate decay parameter

5.1. Experiments on Dynamic Pricing with Positive Lead Time. Firstly, we conduct experiments for perishables' joint ordering and pricing with positive lead time (where $L = 1$). In order to examine the positive impact of dynamic pricing, we consider a fixed-price policy where the agent always takes the fixed best price which achieves the highest revenue. Let MEP and MEP_{FP} be the expected mean epochs profits for the dynamic ordering and pricing policy and the fixed-price ordering policy, respectively, and MDC and MDC_{FP} be the mean epochs disposal cost. After ten thousand simulations, we get the results in Table 3. Table 3 shows that PAQ-DQN achieves better performance than PAQ-A2C when lifetime is

Table 3: Results for dynamic pricing with positive lead time.

Method	Lifetime	MEP	MEP_{FP}	MDC	MDC_{FP}
PAQ-DQN	2	3242.104	2911.139	254.905	291.997
	3	4734.766	4455.814	40.439	110.146
	4	4840.443	4714.712	34.066	28.642
PAQ-A2C	2	3081.546	2305.394	232.000	573.749
	3	4513.786	4474.915	207.470	130.278
	4	4919.332	2278.324	49.802	748.916

2 and 3, but when lifetime is 4, they achieve almost the same results. Mean epochs profits and mean epochs disposal cost for two algorithms are almost all increasing and decreasing with lifetimes, which are in line with expectations, because the longer the lifetime is, the more similar it is to ordinary goods and perishables have more lifetime to sell out under the FIFO policy. From Table 3, it is easy to find out that it is better to adjust the price in a dynamic way so that the price can be adjusted according to the availability of inventory and the remaining life of the product and maximize the profits.

Table 4 shows the comparison between the tabular Q-learning and reinforcement learning methods on mean epoch profits and mean epoch disposal cost. From the table, we can see our proposed PAQ-DQN and PAQ-A2C obviously performs better than the Q-learning method. As we have mentioned before, Q-learning is a tabular method; it stores every state-action value in a table, but in our perishables inventory system, we considered the different ages, so the state space increases exponentially with lifetime, which is inefficient and impractical. Moreover, the amount of computing power and time required increase greatly with lifetime for the Q-learning method.

5.2. Experiments on the Performance of Proposed Algorithms. In this case, we compute the mean epoch profits for the optimal policy and proposed PAQ-DQN and PAQ-A2C with zero lead time. In particular, we set up an upper bound benchmark for this computation and define it as the optimal policy. The optimal policy takes the same price action as the PAQ-DQN and PAQ-A2C in each period, and its order quantity is always equal to the real demand D^t in each period, which means there is always no holding cost, penalty cost, and disposal cost for the planning horizon. Although there may be still some unreasonable place, this can be a useful metric to gauge the performance of the agent. Table 5 shows the computed results after twenty thousand simulations and $MEP_{average}$, $MEP_{average} = (MEP_{optimal} - MEP)/T$, where T denotes the mean difference between the mean epochs profits from deep reinforcement learning methods and the mean epoch profits from the optimal policy. From the table, we can see our proposed algorithms achieve a good performance for three different lifetimes, where the benchmark is a loose upper bound from the real demand D^t and the difference from the average optimal profit is almost always less than the highest possible profit per unit, that is, $MEP_{average} \leq \overline{p} - c$. And the algorithm PAQ-A2C is slightly better than algorithm PAQ-DQN. Figure 1 shows the real epoch profits for the proposed PAQ-DQN and PAQ-A2C (in order to show the variation, we let the initial negative values as zero); from the figure, we can see that two methods quickly reached a relatively flat of profitability and PAQ-A2C showed more stable properties at the beginning of the learning process.

Figures 2–4 show the scatter plots of the profits difference between the optimal policy and the proposed PAQ-DQN algorithm for three different lifetimes. To better show the convergence rate, the figure is drawn on a log-log scale. From three figures, we can see three MEP differences begin to decrease rapidly after about fifty simulations; this demonstrates our deep reinforcement learning method works, the agent gradually learns how to order, and price is near optimal. Besides, the fitting lines in the figures are used to depict the convergence rate, and the following fitting line functions are for lifetime 2, 3, and 4. Here, we also carry out sensitivity analysis to investigate the effects of learning rate α and exploration parameter ε_{decay} for the training of the proposed deep reinforcement learning methods, respectively. Figure 5 demonstrates the MEP for three different learning rates on PAQ-DQN and the learning rate α at 0.001 is the best for three different lifetimes, and α at 0.01 is very close to the best performance. Figure 6 shows the effects of exploration parameters ε_{decay} on PAQ-DQN, and when the exploration parameter ε_{decay} is 1×10^3, the agent gets a higher reward than the other two parameters. And the difference between the three parameters is very obvious. From the above two sensitivity analysis cases, the importance of hyperparameter is verified, and this is a common problem in deep learning. To show the expansibility of the algorithm, we also extend the distribution of the demand into an additive form in Chen et al. [2], where random term has a zero mean. By setting the random term which satisfies a uniform distribution in $[A, B]$, where A and B are symmetric and the absolute value is 2, we get a near-optimal performance with optimal rate 96.344%:

$$\log\left(MEP_{diff}\right) \approx -0.609 \log(\text{epochs}) + 11.794 \left(r^2 = 0.963\right), \tag{16}$$

$$\log\left(MEP_{diff}\right) \approx -0.737 \log(\text{epochs}) + 12.445 \left(r^2 = 0.986\right), \tag{17}$$

$$\log\left(MEP_{diff}\right) \approx -0.575 \log(\text{epochs}) + 11.278 \left(r^2 = 0.956\right). \tag{18}$$

5.3. Experiments on Dynamic Ordering and Pricing with No Fixed Ordering Cost. In this case, when the real epoch profits gradually become *stable* (*stable* means the real epoch profits

Table 4: Results.

Method	lifetime	Lead time	MEP	MDC
PAQ-DQN	2	0	5377.021	18.747
		1	3242.104	254.905
PAQ-A2C	2	0	5385.653	21.222
		1	3081.546	232.000
Q-learning	2	0	4800.622	31.376
		1	592.721	151.462

Table 5: MEP for proposed algorithm and optimal policy.

Method	Lifetime	MEP	$MEP_{optimal}$	$MEP_{average}$	$MEP/MEP_{optimal} * 100\%$
PAQ-DQN	2	5377.021	5691.355	10.477	94.477
	3	5514.710	5691.253	5.884	96.898
	4	5409.587	5664.803	8.507	95.495
PAQ-A2C	2	5385.653	5662.379	9.224	95.113
	3	5590.559	5674.625	2.802	98.519
	4	5428.072	5677.201	8.304	95.612

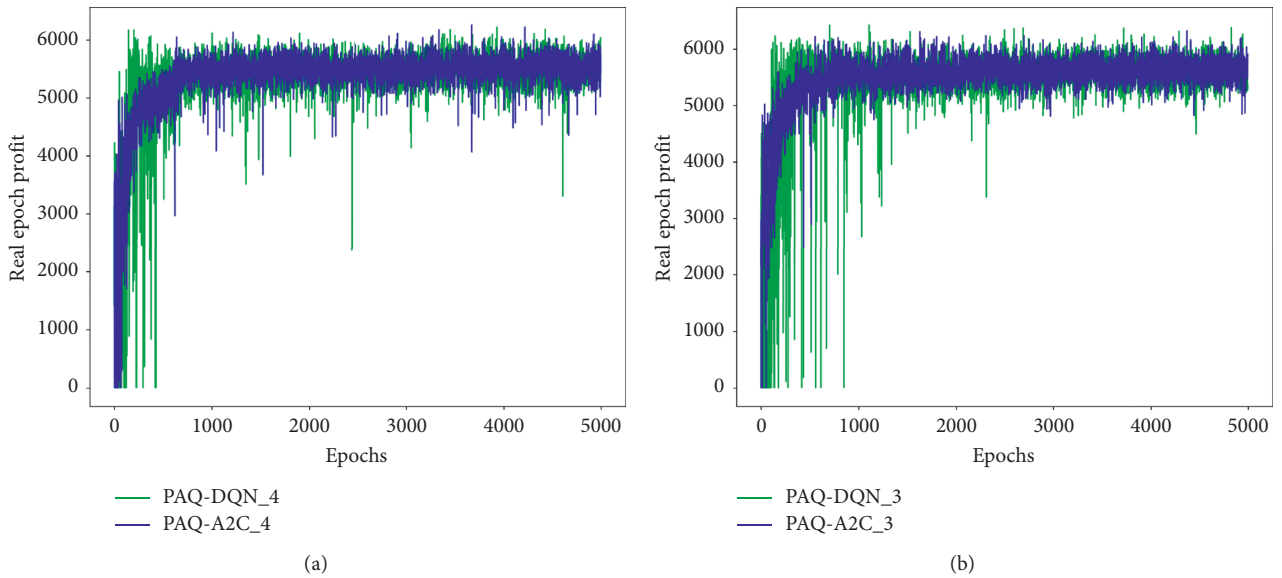


Figure 1: Real epoch profits for PAQ-DQN and PAQ-A2C, where the lifetime 3 and lifetime 4 are shown.

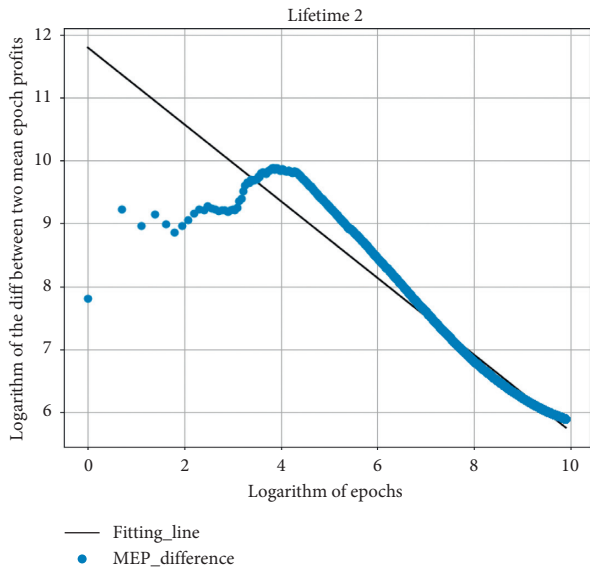


Figure 2: Log-log scale MEP difference.

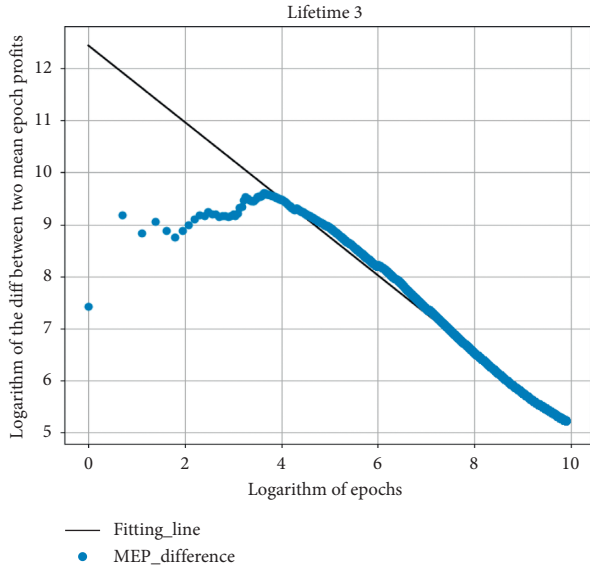


Figure 3: Log-log scale MEP difference.

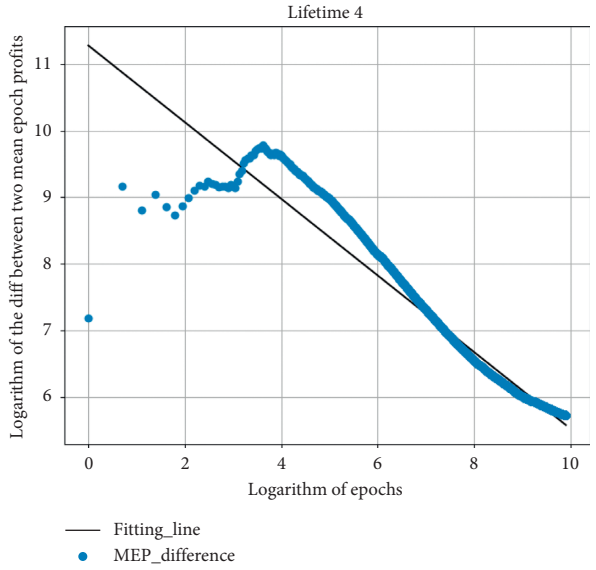


Figure 4: Log-log scale MEP difference.

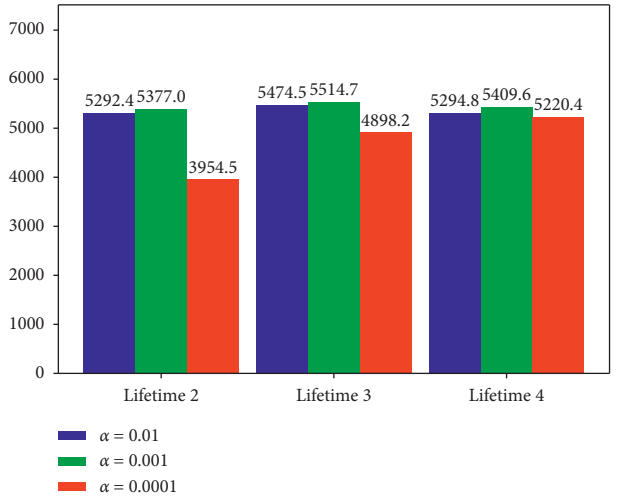


Figure 5: MEP for learning rates α.

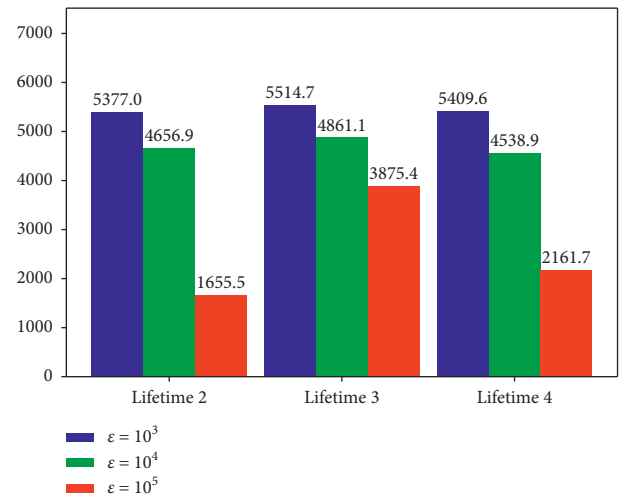


Figure 6: MEP for $\varepsilon_{\text{decay}}$.

always go up and down in a small fixed range over time) with the number of epochs, it indicates that the agent has learned a relatively stable state-to-action mapping relationship. In this section, we first extract the latest n epochs with stable mapping, denoted by dataset $\mathbf{p} = \{[x^{i1}, A^{i1}, \ldots, x^{it}, A^{it}, \ldots, x^{iT}, A^{iT}]_{i=1}^{n}\}$, where x^{it} is the inventory state vector for epoch i and period t. To facilitate discussion, we will use the fragment $\tilde{\mathbf{p}}^t = \{[x^{it}, A^{it}]_{i=1}^{n}\}$ $(t = 1, \ldots, T)$ extracted from $\mathbf{p}$.

Chen et al. [2] has discussed the properties of optimal policies in the joint pricing and inventory system without fixed ordering cost. From the above settings and through our proposed reinforcement learning methods, when there is no fixed ordering cost, we get that the learned order quantity is nonincreasing in both outstanding and on-hand inventory levels. When $L = 0$, the learned price is always equal to the price that achieves highest expected revenue, and when $L > 0$, the learned price is most sensitive to the oldest on-hand inventory.

Figure 7 shows that the order quantity decreases with the inventory position and on-hand inventory. In order to show the sensitivity, we extract the fragment $\tilde{\mathbf{P}}^5$ from $\mathbf{P}$ as an example, where $l = 4, L = 2$, and $n = 1000$. In the inventory state $X^5 = (x_1, x_2, x_3)$, where x_1 is the outstanding order and x_3 is the oldest on-hand inventory, we find that x_1 and x_2 are equal to a fixed value happens more than 500 times out of 1000, and when $x_1 = x_2$, the price decreases with the oldest on-hand inventory. The same results can be obtained from other fragments. In this setting, Figure 8 shows that the price decreases with the oldest on-hand inventory, which means when the oldest inventory increases, the agent tends to set a lower price.

5.4. Experiments on Dynamic Ordering and Pricing with Fixed Ordering Cost. In this part, we will consider the case when there is a fixed ordering cost in this joint pricing and inventory system. We use our propose deep reinforcement learning algorithms to solve this case, and in order to measure the final performance, we set up a loose upper bound as our benchmark. In this benchmark, there are trade-offs between different costs. The price decision is supplied by algorithms. For ease of discussion and simplicity, we assume zero penalty costs and zero disposal costs to be achieved, which mean each demand will be met and each order will be sold within l period. We also assume that the initial inventory is zero; thus, the first order will always be placed at the beginning of the planning horizon. In particular, when $L > 0$, there may be a penalty cost at the beginning. D^t is the real demand in period t, $t = 1, \ldots, l, \ldots, T$. It is obviously unwise to order every period in this setting.

When $L = 0$, taking into account the width of finite lifetime l and the minimization of total cost, it is easy to see that the agent needs to place at least one order every l term and every order is just consumed by the next one. In the first l term, if $[(D^2 + \cdots + D^l) + (D^3 + \cdots + D^l) + \cdots + (D^{l-2} + D^{l-1}) + D^{l-1}] * h \le K$, it only needs one order at the beginning of the first period and ordering quantity $q = D^1 + \cdots + D^l$. Morover, at some point t_o, whether to order depends on the time of last order $t_n, t_o - t_n \le l$, if $[(D^{t_n+1} + \cdots + D^{t_o-1}) + (D^{t_n+2} + \cdots + D^{t_o-1}) + \cdots + D^{t_o-1}] * h \le K$ and $[(D^{t_n+1} + \cdots + D^{t_o}) + (D^{t_n+2} + \cdots + D^{t_o}) + \cdots + D^{t_o}] * h > K$, it should order at point t_o and the order quantity for t_n is $q = D^{t_n} + \cdots + D^{t_o-1}$. When $L > 0$, taking into account the width of finite lifetime l and the minimization of total cost, it is easy to see that the agent needs to place at least one order every l term. In the first l term, there is a penalty cost, $(D^1 + \cdots + D^L) * u$, due to the lag of the order. At some point t_o $(t_o \ne 1)$, whether to order depends on the time of last order t_n and in order to make the subsequent penalty cost zero, $t_o - t_n \le l - L$. If $[(D^{t_n+L+1} + \cdots + D^{t_o+L-1}) + (D^{t_n+L+2} + \cdots + D^{t_o+L-1}) + \cdots + D^{t_o+L-1}] * h \le K$ and $[(D^{t_n+L+1} + \cdots + D^{t_o+L}) + (D^{t_n+L+2} + \cdots + D^{t_o+L}) + \cdots + D^{t_o+L}] * h > K$, it should order at point t_o, and when $t_n = 1$, the order quantity for t_n is $q = D^{t_n} + \cdots + D^{t_o+L-1}$; when $t_n \ne 1$, the order quantity for t_n is $q = D^{t_n+L} + \cdots + D^{t_o+L-1}$.

We consider two different fixed ordering costs K, $K \in \{25, 50\}$, two different penalty costs u, $u \in \{10.78, 4.18\}$ (corresponding, $u/(h+u) \in \{98\%, 95\%\}$), and two different

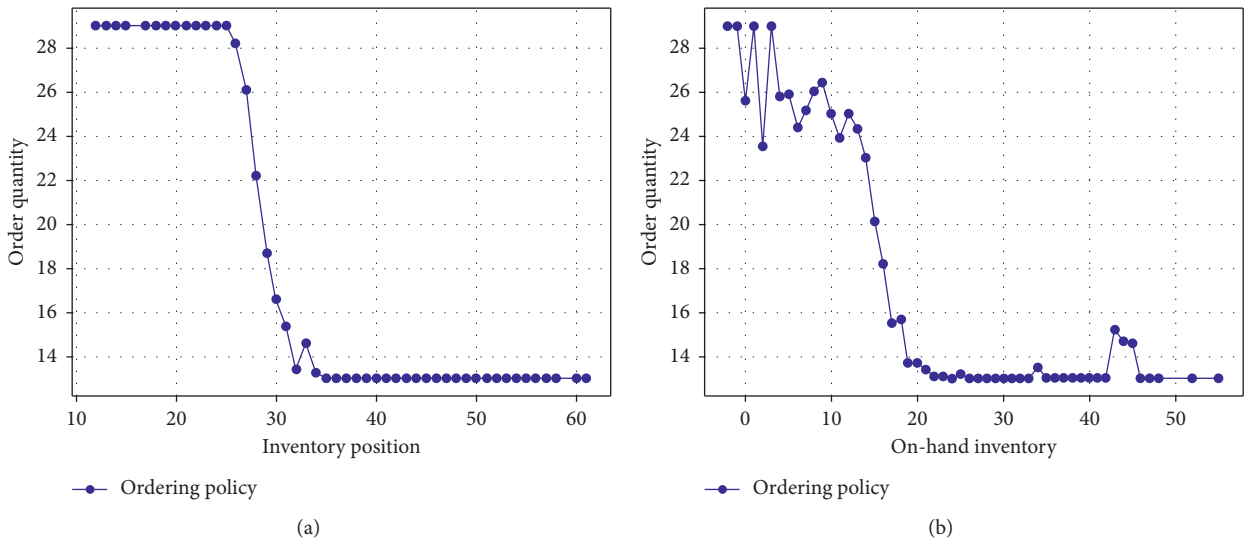


FIGURE 7: This statistic results are from $\tilde{\mathbf{p}}^5$, where $l = 4$ and $L = 2$. When $L = 2$, x^{it} is a vector and order quantity is the mean from the same inventory value.

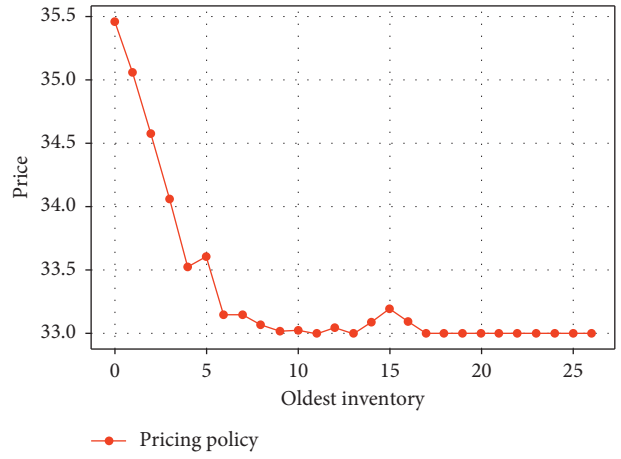


FIGURE 8: This statistic results are from $\tilde{\mathbf{p}}^5$, where $l = 4$ and $L = 2$. When $L = 2$, x^{it} is a vector and price is the mean from the same inventory value.

price-demand functions d^t, and the first one is shown in Table 2 and another one is $d^t = 380 - 10p$. Lifetime $l = 4$, lead time $L = 0, 1$, and a larger order action space is considered for $d^t = 380 - 10p$.

Under all of the above setting, we find that when $L = 0$, the convergent price is always the price that maximizes the expected revenue. This is in line with expectations. Same as the no fixed ordering cost case, the order quantity is nonincreasing in both outstanding and on-hand inventory levels. Table 6 shows the MEP results from PAQ-DQN and PAQ-A2C when the fixed ordering cost is 25 and 50 and price-demand function $d^t = 380 - 10p$ after thirty thousand simulations. From the table, we can see that, under the same conditions, the mean epoch profits MEP decreases with the lead time and the fixed ordering cost. When $L = 0$, algorithm PAQ-A2C performs better than PAQ-DQN, and when $L > 0$, our proposed PAQ-DQN performs better than PAQ-A2C. More interestingly, in our learned convergent policies, we find there exist one or two critical values in the inventory position in each period in each case when $L = 0$ and $L = 1$. We denote cv^t as the critical value in each period for the one critical value cases and cv_1^t and cv_2^t in each period for the two critical value cases, $cv_1^t < cv_2^t$. In the one critical value case, when $x^t < cv^t$, there will be a fixed order quantity q_1; when $x^t \geq cv^t$, the fixed order quantity is q_2. In the case of two critical values, when $x^t < cv_1^t$, the fixed order quantity is q_1; when $cv_1^t \leq x^t < cv_2^t$, the fixed order quantity is q_2; when $x^t \geq cv_2^t$, the fixed order quantity is q_3. For more details about obtaining the critical values, see Appendix.

Table 7 shows the MEP of learned policies from algorithm PAQ-DQN; from the table, we can see our learned policies achieve a higher optimal rate and are closer to the upper bound, compared to Table 6. Table 8 shows the MEP comparison between the learned policies and the algorithm PAQ-DQN; from the table, we can see our learned policies achieve a higher MEP and lower MDC, which means our learned policies are working well.

Table 6: MEP with fixed ordering cost.

Method	Lifetime	Lead time	K	$u/h+u$ (%)	MEP	$\text{MEP}_{\text{benchmark}}$	$\text{MEP}/\text{MEP}_{\text{benchmark}} * 100\%$
PAQ-DQN	4	0	50	98	15041.55	16172.39	93.01
		1	50	98	14484.18	15576.99	92.98
		0	25	98	15753.94	16507.79	95.43
		1	25	98	14544.00	15436.53	94.22
PAQ-A2C	4	0	50	98	15116.77	16178.22	93.44
		1	50	98	14223.13	15423.03	92.22
		0	25	98	15813.49	16502.16	95.83
		1	25	98	14896.70	15858.32	93.94

Table 7: MEP for the learned policies.

Lifetime	Lead time	K	$\text{MEP}_{\text{policies}}$	$\text{MEP}_{\text{benchmark}}$	$\text{MEP}_{\text{policies}}/\text{MEP}_{\text{benchmark}} * 100\%$
4	0	25	16013.56	16525.75	96.90
	1	25	14911.06	15540.62	95.95
4	0	50	15421.45	16197.45	94.10
	1	50	14771.21	15583.89	94.79

Table 8: Comparison for PAQ-DQN.

Lifetime	Lead time	K	MEP	$\text{MEP}_{\text{policies}}$	MDC	$\text{MDC}_{\text{policies}}$
4	0	25	15753.94	16013.56	2.45	0
	1	25	14544.00	14911.06	36.87	0
4	0	50	15041.55	15241.45	1.92	0
	1	50	14484.18	14771.21	97.67	56.97

Table 9: MEP for new state forms.

Lifetime	Lead time	K	$u/h+u$ (%)	$\text{MEP}_{\text{PAQ-DQN}}$	$\text{MEP}_{\text{PAQ-A2C}}$	$\text{MEP}_{\text{new-state}}$
3	0	50	98	15052.747	15126.553	15465.963
	1	50	98	13685.987	13544.115	13842.541
	2	50	98	8310.238	7018.234	9633.665
3	0	50	95	14937.285	14996.444	15604.507
	1	50	95	13894.832	13847.474	14317.453
	2	50	95	10470.134	9989.935	10299.061

Table 10: MEP for new state forms.

Lifetime	Lead time	K	$u/h+u$ (%)	$\text{MEP}_{\text{PAQ-DQN}}$	$\text{MEP}_{\text{PAQ-A2C}}$	MEP_{S^1}	MEP_{S^2}
3	0	50	95	14937.285	14996.444	15604.507	15309.898
			98	15052.747	15126.553	15465.963	15117.522

Table 11: Results for dynamic pricing with positive lead time.

Method	Lifetime	MEP	MEP_{FP}	MDC	MDC_{FP}
PAQ-DQN	2	2761.297	2648.036	371.397	371.135
	3	4832.157	4624.937	51.167	73.137
	4	4917.960	4787.649	28.737	42.665
PAQ-A2C	2	2739.912	2492.352	330.639	329.539
	3	4647.317	4490.300	94.006	102.618
	4	4807.079	4638.502	84.877	84.269

The above discussion is mainly based on the current inventory state. Next, we will try to add the historical inventory states and action information to the state to discuss its impact on the final performance. Here, we define the new state to be $S^t = (X^{t-L}, A^{t-L}, X^{t-L+1}, \ldots, X^t)$ when $L > 0$, and when $L = 0$, we also discuss the gradual influence of the addition of information in the state on the final performance. At the same time the dimension of the state accompanying the increase in information will also increase.

Table 9 shows the results after ten thousand simulations and there we add the inventory state and action from the previous period for the new state when $L = 0$. From the table,

Table 12: Results.

Method	lifetime	Lead time	MEP	MDC
PAQ-DQN	2	0	5394.735	21.519
		1	2761.297	371.397
PAQ-A2C	2	0	5367.027	16.139
		1	2739.912	330.639
Q-learning	2	0	4437.220	87.911
		1	<0	0

Table 13: MEP for proposed algorithm and optimal policy.

Method	Lifetime	MEP	$MEP_{optimal}$	$MEP_{average}$	$MEP/MEP_{optimal} * 100\%$
PAQ-DQN	2	5349.735	5670.955	10.707	94.336
	3	5574.282	5693.097	3.960	97.913
	4	5435.068	5685.511	8.348	95.595
PAQ-A2C	2	5367.027	5695.111	10.936	94.239
	3	5627.395	5686.337	1.964	98.963
	4	5504.487	5694.402	6.330	96.665

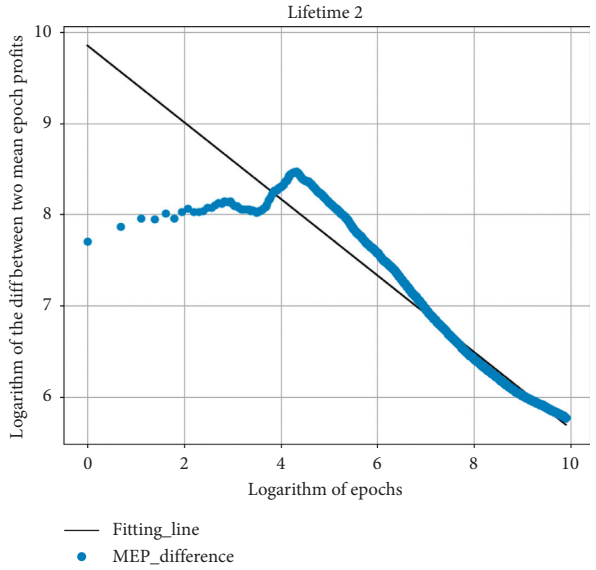


Figure 9: Log-log scale MEP difference for lifetime 2.

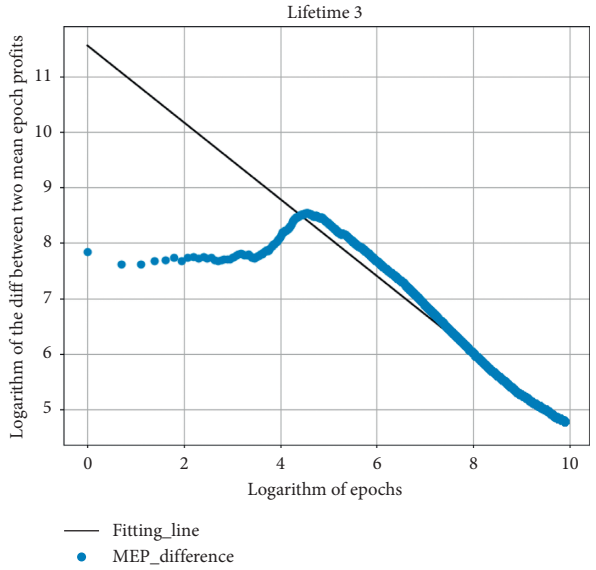


Figure 10: Log-log scale MEP difference for lifetime 3.

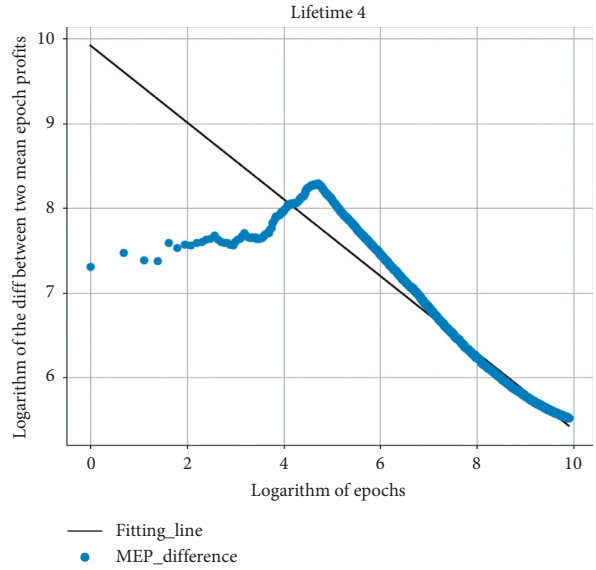


Figure 11: Log-log scale MEP difference for lifetime 4.

Table 14: Critical values for $L = 0$.

(K, L)	cv	1	2	3	4	5	6	7	8	9	10	11	12
(25, 0)	cv^1	−6	−6	−5	−6	−6	−6	−6	−7	−6	−6	−5	−5
	cv^2	22	23	23	24	23	22	24	23	24	23	23	24
		13	14	15	16	17	18	19	20	21	22	23	24
	cv^1	−6	−6	−6	−6	−5	−5	−6	−6	−5	−5	−7	−6
	cv^2	24	23	23	24	23	22	22	23	23	24	23	24
		25	26	27	28	29							
	cv^1	−5	−6	−5	−4	−5							
	cv^2	23	24	23	24	23							
(K, L)	cv	1	2	3	4	5	6	7	8	9	10	11	12
(50,0)	cv^1	−9	−10	−10	−11	−10	−9	−10	−10	−11	−10	−9	−11
	cv^2	12	13	13	13	14	13	12	13	13	13	13	13
		13	14	15	16	17	18	19	20	21	22	23	24
	cv^1	−10	−10	−10	−11	−10	−11	−10	−11	−9	−11	−11	−8
	cv^2	13	13	13	13	13	13	13	12	13	13	13	13
		25	26	27	28	29							
	cv^1	−9	−13	−10	−10	−10							
	cv^2	13	13	13	13	12							

Table 15: Critical values for $L = 1$.

(K, L)	cv	1	2	3	4	5	6	7	8	9	10	11	12
(25, 1)	cv^1	121	100	121	100	121	103	121	102	121	106	121	105
		13	14	15	16	17	18	19	20	21	22	23	24
	cv^1	121	108	121	110	121	114	121	116	121	116	121	118
		25	26	27	28	29							
	cv^1	121	121	121	121	121							
(K, L)	cv	1	2	3	4	5	6	7	8	9	10	11	12
(50, 1)	cv^1	62	62	63	63	64	65	63	64	63	64	64	64
		13	14	15	16	17	18	19	20	21	22	23	24
	cv^1	64	64	65	64	63	64	64	63	64	64	64	64
		25	26	27	28	29							
	cv^1	64	63	64	64	64							

Table 16: MEP with fixed ordering cost.

Lifetime	Lead time	K	$u/h+u$ (%)	MEP	$\text{MEP}_{\text{benchmark}}$	$\text{MEP}/\text{MEP}_{\text{benchmark}} * 100\%$
4	0	50	98	15378.025	16127.167	95.355
	1	50	98	14484.18	15576.99	92.98
4	0	25	98	15753.94	16507.79	95.43
	1	25	98	14844.676	15622.779	95.019

we can see that the new state contains more information almost all performs better than the single current inventory state, which means the current decisions of the agent are influenced by not only the current inventory state but also the inventory states of the previous periods. In Table 10, MEP_{S^1} and MEP_{S^2}, respectively, represent the inventory state and action information of the previous period and the previous two periods added to the current state. From the table, we find that the final performance did not get better and better with the continuous addition of the historical information, which also confirms that the dimensions of the state mentioned above continue to increase with the addition of information, which may have a negative impact on learning.

6. Conclusions

In this paper, we investigate a joint pricing and inventory control problem and obtain near-optimal pricing and replenishment policies for stochastic perishable inventory systems with positive lead time by deep reinforcement learning algorithms. Through our designed algorithms, we show that, in a perishable inventory control problem, the expected profit is maximized by adjusting the price according to the availability of inventory and the remaining lives of the items. We consider the case of no fixed ordering cost and the one involving a fixed ordering cost and find near-optimal policies for both cases. Our findings when a fixed ordering cost is involved contribute to the literature of inventory control for perishables, which has not been studied before. In this paper, we only focus on a single agent's joint pricing and inventory control problem. However, multiple agents are usually involved in supply chains, and their interactions may have a big impact on each agent's pricing and inventory decisions. Therefore, the study of the competition and cooperation of participants under complete and incomplete information is an interesting topic for future research.

Appendix

This section is for the discussion about lost-sales case in Section 5.

A. Experiment on Dynamic Pricing with Positive Lead Time

Table 11 shows that it is better to adjust the price in a dynamic way so that the price can be adjusted according to the availability of inventory and the remaining life of the product and maximize the profits. From Table 12, we can also see our proposed method PAQ-DQN achieves better performance than PAQ-A2C when lead time is positive and our proposed deep reinforcement learning methods obviously perform better than the Q-learning method.

B. Experiments on the Performance of Proposed Algorithms

In this case, we compute the mean epoch profits for the optimal policy and proposed PAQ-DQN and PAQ-A2C with zero lead time. Table 13 shows the computed results after twenty thousand simulations. From the table, we can see our proposed algorithms achieve a good performance for three different lifetimes, where the benchmark is a loose upper bound from the real demand D^t and the difference from the average optimal profit is almost always less than the highest possible profit per unit. And the algorithm PAQ-A2C is slightly better than algorithm PAQ-DQN. Figures 9–11 show the scatter plots of the profits' difference between the optimal policy and the proposed PAQ-DQN algorithm for three different lifetimes. To better show the convergence rate, the figure is drawn on a log-log scale. From three figures, we can see three MEP differences all begin to decrease rapidly after about ninety simulations; this demonstrates our deep reinforcement learning method works, the agent gradually learns how to order, and price is optimal. Besides, the fitting lines in the figures are used to depict the convergence rate, and the following fitting line functions are for lifetime 2, 3, and 4:

$$\log\left(\text{MEP}_{\text{diff}}\right) \approx -0.401\log(\text{epochs}) + 9.862\left(r^2 = 0.960\right), \tag{B.1}$$

$$\log\left(\text{MEP}_{\text{diff}}\right) \approx -0.677\log(\text{epochs}) + 11.531\left(r^2 = 0.982\right), \tag{B.2}$$

$$\log\left(\text{MEP}_{\text{diff}}\right) \approx -0.429\log(\text{epochs}) + 9.973\left(r^2 = 0.960\right). \tag{B.3}$$

B.1. Experiments on Dynamic Ordering and Pricing with no Fixed Ordering Cost. In this section, under the same settings as the backlogging case, we get the same results where the learned order quantity is nonincreasing in both outstanding and on-hand inventory levels. When $L = 0$, the learned price is always equal to the price that achieves highest expected revenue, and when $L > 0$, the learned price is most sensitive to the oldest on-hand inventory.

B.2. Experiments on Dynamic Ordering and Pricing with Fixed Ordering Cost. Firstly, we introduce the steps to get the critical values mentioned in the backlogging case. We first

extract the latest n epochs with stable mapping, denote by dataset $\mathbf{p} = \left\{[x^{i1}, A^{i1}, \ldots, x^{it}, A^{it}, \ldots, x^{iT}, A^{iT}]_{i=1}^{n}\right\}$, where x^{it} is the inventory state vector for epoch i and period t. To facilitate discussion, we will use the fragment $\widetilde{\mathbf{p}}^t = \{[x^{it}, A^{it}]_{i=1}^{n}\}$ $(t = 1, \ldots, T)$ extracted from $\widetilde{\mathbf{p}}$. Based on the fragment $\widetilde{\mathbf{p}}^t = \{[x^{it}, A^{it}]_{i=1}^{n}\}$ $(t = 1, \ldots, T)$, we can observe the relationship between the inventory level and the order quantity and price. In the backlogging case, when $l = 4$, $L \in [0, 1]$, $K \in [25, 50]$, and $u/h + u = 98\%$, we get the following ordering and pricing policies. In each epoch, we set the first period as a zero initial inventory, so we cannot observe the critical values, and the following values are obtained from the second period. Tables 14 and 15 show the critical values in the different settings. When $L = 0$ and $K = 25$, the price is always 32 and $q_1 = 80, q_2 = 65$, and $q_3 = 35$. When $L = 0$ and $K = 50$, the price is always 32 and $q_1 = 95, q_2 = 65$, and $q_3 = 50$; when $L = 1$ and $K = 50$, the price is 32 except for the first period and $q_1 = 120$ and $q_2 = 0$; when $L = 1$ and $K = 25$, when inventory position is less than the critic value cv^1, the price is 32; otherwise, the price is 33, $q_1 = 75$, and $q_2 = 45$. In the lost-sales case, Table 16 shows the MEP for the different settings. And the same learned convergent policies structure as backlogging case can be obtained from this lost-sales case.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] I. Z. Karaesmen, A. Scheller-Wolf, and B. Deniz, "Managing perishable and aging inventories: review and future research directions," in *Planning Production and Inventories in the Extended Enterprise*, pp. 393–436, Springer, Berlin, Germany, 2011.

[2] X. Chen, Z. Pang, and L. Pan, "Coordinating inventory control and pricing strategies for perishable products," *Operations Research*, vol. 62, no. 2, pp. 284–300, 2014.

[3] S. K. Chaharsooghi, J. Heydari, and S. H. Zegordi, "A reinforcement learning model for supply chain ordering management: an application to the beer game," *Decision Support Systems*, vol. 45, no. 4, pp. 949–959, 2008.

[4] I. Dogan and A. R. Güner, "A reinforcement learning approach to competitive ordering and pricing problem," *Expert Systems*, vol. 32, no. 1, pp. 39–48, 2015.

[5] A. Kara and I. Dogan, "Reinforcement learning approaches for specifying ordering policies of perishable inventory systems," *Expert Systems with Applications*, vol. 91, pp. 150–158, 2018.

[6] J. Ke, F. Xiao, H. Yang, and J. Ye, "Optimizing online matching for ride-sourcing services with multi-agent deep reinforcement learning," 2019, http://arxiv.org/abs/1902.06228.

[7] S. A. M. Shihab, C. Logemann, D.-G. Thomas, and P. Wei, "Autonomous airline revenue management: a deep reinforcement learning approach to seat inventory control and overbooking," 2019, http://arxiv.org/abs/1902.06824.

[8] E. Presman and S. P. Sethi, "Inventory models with continuous and Poisson demands and discounted and average costs," *Production and Operations Management*, vol. 15, no. 2, pp. 279–293, 2006.

[9] O. Caliskan-Demirag, Y. Chen, and Y. Yang, "Ordering policies for periodic-review inventory systems with quantity-dependent fixed costs," *Operations Research*, vol. 60, no. 4, pp. 785–796, 2012.

[10] O. Alp, W. Tim Huh, and T. Tan, "Inventory control with multiple setup costs," *Manufacturing & Service Operations Management*, vol. 16, no. 1, pp. 89–103, 2013.

[11] A. T. Almaktoom, "Stochastic reliability measurement and design optimization of an inventory management system," *Complexity*, vol. 2017, Article ID 1460163, 9 pages, 2017.

[12] R. Azghandi, J. Griffin, and M. S. Jalali, "Minimization of drug shortages in pharmaceutical supply chains: asimulation based analysis of drug recall patterns and inventory policies," *Complexity*, vol. 2018, Article ID 6348413, 14 pages, 2018.

[13] C. Li, H. Guo, Y. Zhang, S. Deng, and Y. Wang, "An improved differential evolution algorithm for a multicommodity location-inventory problem with false failure returns," *Complexity*, vol. 2018, Article ID 1967398, 2018.

[14] X. Gan, S. P. Sethi, and L. Xu, "Simultaneous optimization of contingent and advance purchase orders with fixed ordering costs," *Omega*, vol. 89, pp. 227–241, 2019.

[15] S. Nahmias and W. P. Pierskalla, "Optimal ordering policies for a product that perishes in two periods subject to stochastic demand," *Naval Research Logistics Quarterly*, vol. 20, no. 2, pp. 207–229, 1973.

[16] S. Nahmias, "Optimal ordering policies for perishable inventory-II," *Operations Research*, vol. 23, no. 4, pp. 735–749, 1975.

[17] B. E. Fries, "Optimal ordering policy for a perishable commodity with fixed lifetime," *Operations Research*, vol. 23, no. 1, pp. 46–61, 1975.

[18] S. Nahmias, "Perishable inventory theory: a review," *Operations Research*, vol. 30, no. 4, pp. 680–708, 1982.

[19] G. P. Prastacos, "Blood inventory management: an overview of theory and practice," *Management Science*, vol. 30, no. 7, pp. 777–800, 1984.

[20] M. E. Ferguson and O. Koenigsberg, "How should a firm manage deteriorating inventory?," *Production and Operations Management*, vol. 16, no. 3, pp. 306–321, 2007.

[21] Q. Chen, Q. Xu, and W. Wang, "Optimal policies for the pricing and replenishment of fashion apparel considering the effect of fashion level," *Complexity*, vol. 2019, Article ID 9253605, 12 pages, 2019.

[22] Y. Li, A. Lim, and B. Rodrigues, "Note-pricing and inventory control for a perishable product," *Manufacturing & Service Operations Management*, vol. 11, no. 3, pp. 538–542, 2009.

[23] C. Li and M. Lu, "Joint price and inventory optimization under minimax regret," *SSRN Electronic Journal*, 2017.

[24] Y. Li, B. Cheang, and A. Lim, "Grocery perishables management," *Production and Operations Management*, vol. 21, no. 3, pp. 504–517, 2012.

[25] S. Bhattacharyya, V. Snasel, A. Hassanien, S. Saha, and B. Tripathy, *Deep Learning: Research and Applications, De Gruyter Frontiers in Computational Intelligence*, De Gruyter, Berlin, Germany, 2020, https://books.google.com/books?id=yEj2DwAAQBAJ.

[26] I. Giannoccaro and P. Pontrandolfo, "Inventory management in supply chains: a reinforcement learning approach,"

International Journal of Production Economics, vol. 78, no. 2, pp. 153–161, 2002.

[27] C. Jiang and Z. Sheng, "Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6520–6526, 2009.

[28] Z. Sui, A. Gosavi, and L. Lin, "A reinforcement learning approach for inventory replenishment in vendor-managed inventory systems with consignment inventory," *Engineering Management Journal*, vol. 22, no. 4, pp. 44–53, 2010.

[29] M. H. F. Zarandi, S. V. Moosavi, and M. Zarinbal, "A fuzzy reinforcement learning algorithm for inventory control in supply chains," *International Journal of Advanced Manufacturing Technology*, vol. 65, no. 1–4, pp. 557–569, 2013.

[30] R. Rana and F. S. Oliveira, "Dynamic pricing policies for interdependent perishable products or services using reinforcement learning," *Expert Systems with Applications*, vol. 42, no. 1, pp. 426–436, 2015.

[31] C. J. C. H. Watkins, *Learning from delayed rewards*, Ph.D. thesis, 1989.

[32] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[33] C. Darken, J. Chang, and J. Moody, "Learning rate schedules for faster stochastic gradient search," in *Proceedings of the 1992 IEEE Workshop on Neural Networks for Signal Processing II*, pp. 3–12, IEEE, Helsingoer, Denmark, 1992.

Hindawi
Complexity
Volume 2021, Article ID 6675279, 15 pages
https://doi.org/10.1155/2021/6675279

Research Article

Handling Imbalance Classification Virtual Screening Big Data Using Machine Learning Algorithms

Sahar K. Hussin,[1] **Salah M. Abdelmageid**,[2] **Adel Alkhalil**,[3] **Yasser M. Omar**,[4] **Mahmoud I. Marie**,[5] **and Rabie A. Ramadan**[3,6]

[1]*Communication and Computers Engineering Department Alshrouck Academy, Cairo, Egypt*
[2]*Computer Engineering Department, Collage of Comp. Science and Engineering, Taibah University, Medina, Saudi Arabia*
[3]*College of Computer Science and Engineering, University of Hai'l, Hai'l, Saudi Arabia*
[4]*Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt*
[5]*Computer and System Engineering Department, Al-Azhar University, Cairo, Egypt*
[6]*Computer Engineering Department, Cairo Universality, Cairo, Egypt*

Correspondence should be addressed to Rabie A. Ramadan; rabie@rabieramadan.org

Received 28 November 2020; Revised 19 December 2020; Accepted 2 January 2021; Published 28 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

Virtual screening is the most critical process in drug discovery, and it relies on machine learning to facilitate the screening process. It enables the discovery of molecules that bind to a specific protein to form a drug. Despite its benefits, virtual screening generates enormous data and suffers from drawbacks such as high dimensions and imbalance. This paper tackles data imbalance and aims to improve virtual screening accuracy, especially for a minority dataset. For a dataset identified without considering the data's imbalanced nature, most classification methods tend to have high predictive accuracy for the majority category. However, the accuracy was significantly poor for the minority category. The paper proposes a *K*-mean algorithm coupled with Synthetic Minority Oversampling Technique (SMOTE) to overcome the problem of imbalanced datasets. The proposed algorithm is named as KSMOTE. Using KSMOTE, minority data can be identified at high accuracy and can be detected at high precision. A large set of experiments were implemented on Apache Spark using numeric PaDEL and fingerprint descriptors. The proposed solution was compared to both no-sampling method and SMOTE on the same datasets. Experimental results showed that the proposed solution outperformed other methods.

1. Introduction

The discovery of new medication to cure human illnesses is progressively hard, expensive, and tedious [1]. A wide variety of atoms and molecules must be chosen and prepared to generate a set of predetermined drugs. The drug discovery process can take between 12 and 15 years, with a possibility of failure, and expenses are worth more than one billion dollars [2]. Virtual screening is the most critical process in drug discovery. It is employed to search for small chemical compounds (molecules) in libraries to identify structures that have an affinity to bind to a drug target or protein receptor [3]. Up to 1010 libraries of virtual screening exist, and since this record keeps increasing, traditional classification methods have become insufficient to manage such large amounts of datasets [4]. One of the well-known repositories is PubChem [5] for small molecules and their biological properties. It offers several resources that are unfortunately constrained by the unbalanced nature of high-throughput screening (HTS) data. These data usually contain a few hundred active compounds, excluding many inactive compounds.

Dataset imbalances occur when one of the classes is described by a minimal number of samples, typically of major importance, compared to the other classes [6]. This problem may distort prediction accuracy in the used

classification models, which leads to poor classification performance. In HTS experiments, thousands of compounds are usually screened; nevertheless, a small fraction of the tested compounds are classified to be active while other classes are recognized inactive [7]. Such data imbalance affects the accuracy and precision of activity predictions in individual virtual screening datasets. Using the binary classification of an imbalanced dataset, an instance of one class became fewer compared to another class. The minority class is known as the class with fewer cases, and the other is called the majority class [8].

Current classification models such as k-nearest neighbor (KNN), random forest (RF), multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), logistic regression (LG), and gradient boosting (GBT) depend on a sufficient, representative, and reasonably balanced collection of training data to draw an approximate boundary for decision-making between different groups. These learning algorithms are utilized in a variety of fields, including financial forecasting and text classification [9]. Despite current advances in machine learning (ML), developing successful algorithms that learn from unbalanced datasets remains a daunting task. In ML, many approaches have been created to deal with imbalanced data. However, very few algorithms have been able to handle problems related to negatives and false positives. Positive or negative states usually dominate the unbalanced dataset. Therefore, specificity and recall (sensitivity) are very vital when processing an imbalanced dataset [10]. The increase in sensitivity increases the true-positive expectations of the model and reduces false negatives. Likewise, an upgrade in specificity increases true-negative expectations and thus reduces wrong responses. Therefore, it is critical that, for a good model, the gap between sensitivity and specificity metrics should be as small as possible [11].

Although some of the researchers highlighted the problem of negatives and false positives when using PubChem data, to the best of our knowledge, no technique has been reported to address the problem effectively. It was also reported that the imbalance problem obstructed the classification accuracy of bioactivity [12]. In another study [13], the authors compared the performance of seven different descriptive classifiers based on their ability to deal with unbalanced datasets. Another research stated that multilevel SVM-based algorithms outperformed certain algorithms such as (1) traditional SVM, (2) weighted SVM, (3) neural networks, (4) linear regression, (5) Naïve Bayes (NB), and (6) C4.5 tree with imbalanced groups, missing values, and real health-related data [14].

Besides, the main concept of resampling is to reduce variance between class samples by preprocessing the training data. In other words, the resampling approach is used in training samples in order to achieve the same number of samples for each class to adjust the previous majority and minority sample distributions. There are two basic methods in the traditional resampling methodology, namely, undersampling and oversampling [15]. Undersampling produces a smaller number of majority samples while all minority samples are retained. The predominant class samples will be eliminated randomly before a satisfactory ratio has been accomplished for all groups. Undersampling is ideal for applications where there is an enormous number of majority samples, and the reduction of training samples would minimize the training time for the model. However, a drawback with undersampling that it discards samples contributes to the loss of majority class information [16].

Another approach to address the imbalanced data is oversampling. By replicating samples, it raises the number of samples in the minority groups [17]. The benefit from oversampling is that because all samples are used, no knowledge is lost. Oversampling, however, has its own drawback. It contributes to higher processing costs by generating extra training samples. Therefore, to compensate for that limitation, more efficient algorithms are needed.

Although resampling methods are usually used to solve problems with imbalances in the class, there is little defined strategy to identify the acceptable class distribution for a particular dataset [18]. As a result, the optimal class distribution differs from one dataset to another. Recent variants of resampling approaches derived from oversampling and undersampling overcome some of the shortcomings of current technologies, including SMOTE (Synthetic Minority Oversampling Techniques). SMOTE is one of the most important oversampling approaches that generate interpolation instances, which is added to the training samples without duplicating the samples in the class of the minority. The SMOTE approach examines the KNN of the minority class test that will be utilized as a base for the new synthetic sample [19]. If created instances are smaller than the size of the initial dataset, the approach randomly selects the original instances utilized to create the artificial ones. If instances are larger than the size of the original dataset, the approach iterates over the dataset, creating an artificial instance per original instance until it reaches the previous scenario [20]. SMOTE is considered as an oversampling technique that produces synthetic minority class samples. This is theoretically performing better than simple oversampling, and it is commonly used. For example, SMOTE was utilized to detect network intrusions [21] or speech boundary sentence, to predict species distribution [22]. SMOTE will be utilized in this research.

Data mining techniques can help to reduce promising candidate chemicals for interaction with specific molecular targets before they are experimentally evaluated [23]. In theory, this can help to speed up the drug development process. However, the improvement of accurate prediction models for HTS is difficult. For datasets such as those taken from HTS experiments, the achievement of high predictability accuracy may be misleading since this may be accompanied by an unacceptable false-positive rate [24] as high accuracy does not always imply a small proportion of false predictions.

In the event of a large class imbalance, this paper attempts to address the most effective variant of data preprocessing to enhance data imbalance accuracy, which favors the collection of interactions that increase the overall accuracy of a learning model. We propose a SMOTE coupled with k-mean method to classify several imbalanced

PubChem datasets with the goal of (1) validating whether *k*-mean with SMOTE affects the output of established models and (2) exploring if KSMOTE is appropriate and useful in finding interesting samples from broad datasets. Our model is also applied to different ML algorithms (random forest, decision tree, multilayer perceptron, logistic regression, and gradient boosting) for comparison purposes with three different datasets. The paper also introduces a procedure for data sampling to increase the sensitivity and specificity of predicting several molecules' behavior. The proposed approach is implemented on standalone clusters for Apache Spark 2.4.3 in order to address the imbalance in a big dataset.

The remainder of this paper is structured as follows. Section 2 provides a general description of class imbalance learning concepts and reviews the related research conducted on the subject matter. Section 3 explains how the proposed approach for VS in drug discovery was developed. Section 4 presents performance evaluations and experimental results. Section 5 presents the discussion of our proposal. Finally, Section 6 highlights the conclusions and topics of study for future research.

2. Related Work

For the paper to be self-contained, this section reviews the most related work to the VS research, techniques, problems, and state-of-the-art solutions. It also examines some of the big data frameworks that could help solve the problem of imbalanced datasets.

Since we live in the technological era where older storage and processing technologies are not enough, computing technologies must be scaled to handle a massive amount of data generated by different devices. The biggest challenge in handling such volumes of data is the speed at which they will grow much faster than the computer resources. One of the research areas that generate huge data to be analyzed is searching and discovering medicines. The proposed methods aim to find a molecule capable of binding and activating or inhibiting a molecular target. The discovery of new drugs for human diseases is exceptionally complicated, expensive, and time-consuming. Drug discovery uses various methods [25] based on a statistical approach to scan for small molecule libraries and determines the structures most likely to bind to a compound. However, the drug target is a protein receptor that is involved in a metabolic cycle or signaling pathway by which a particular disease disorder is established or another anatomy.

There are two VS approaches, which are ligand-based VS (LBVS) and structure-based virtual screening (SBVS) [26]. LBVS depends on the existing information about the ligands. It utilizes the knowledge about a set of ligands known to be active for the given drug target. This type of VS uses the mining of big data analytics. Training binary classifiers by a small part of a ligand can be employed, and very large sets of ligands can be easily classified into two classes: active and nonactive ones. SBVS, on the other side, is utilized to dock experimentally. However, 3D protein structural information is required [27], as shown in Figure 1.

K-mean clustering is one of the simplest nonsupervised learning algorithms, which was first proposed by Macqueen in 1967. It has been applied by many researchers to solve some of the problems of known groups [28]. This technique classifies a particular dataset into a certain number of groups. The algorithm randomly initializes the center of the groups. Then, it calculates the distance between an object and the midpoint of each group. Next, each data instance is linked to the nearest center, and the cluster centers are recalculated. The distance between the center and each sample is calculated by the following equation:

$$\text{Euclidean distance} = \sum_{i=1}^{c} \sum_{j=1}^{ci} \|Xi - Yj\|, \tag{1}$$

where the Euclidean distance between the data point Xi and cluster center y is d, Ci is the total number of data points i in cluster, and c is the total number of cluster centers. All of the training samples are first grouped into K groups (the experiment with diverse K values runs to observe the result). Suitable training samples from the derived clusters are selected. The key idea is that there are different groups in a dataset, and each group appears to have distinct characteristics. When a cluster includes samples of large majority class and samples of low minority class, it functions as a majority class sample. If on the other side, a cluster has extra minority samples and fewer majority samples, it acts more like a minority class. Therefore, by considering the number of majority class samples to that of minority class samples in various clusters, this method oversamples the required number of minority class samples from each cluster.

Several approaches have been proposed in the literature to handle big data classification including classification algorithms, random forest, decision tree, multilayer perceptron, logistic regression, and gradient boosting.

Classification algorithms (CA) are mainly depending on machine learning (ML) algorithms, where they play a vital role in VS for drug discovery. It can be considered as an LBVS approach. Researchers widely used the ML approach to create a binary classification model that is a form of filter to classify ligands as active or inactive in relation to a particular protein target. These strategies need fewer computational resources, and because of their ability to generalize, they find more complex hits than other earlier approaches. Based on our experience, we believe that many classification algorithms can be utilized for dealing with unbalanced datasets in VS, such as SVM, RF, Naïve Bayes, MLP, LG, ANN, DT, and GBT. Five ML algorithms are applied in this paper RF, DT, MLP, LG, and GBT [29].

Random forest (RF) is an ensemble learning approach in which multiple decision trees are constructed based on training data and a majority voting mechanism. Like KNN, it is utilized to predict classification or regression for new inputs. The key advantage of RF is that it can be utilized for problems that need classification and regression. Besides, RF is the ability to manage many higher-dimensional datasets. It has a powerful strategy for determining the lack of information and preserving accuracy when much of the information is missing [29].

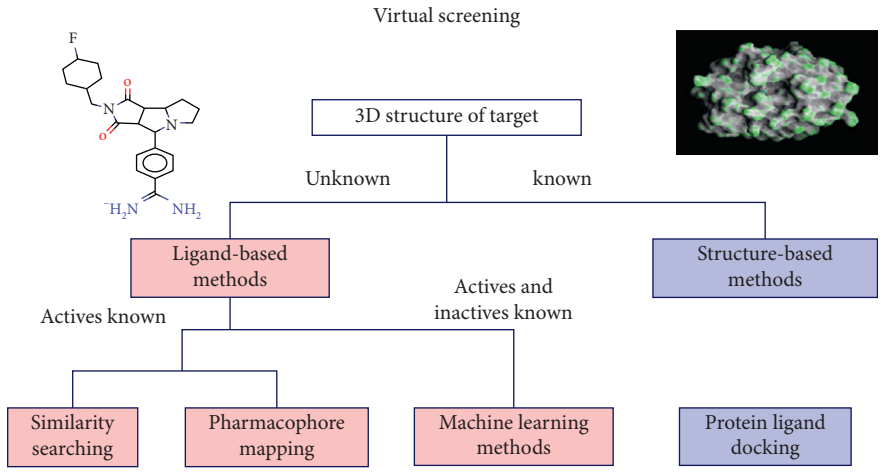


Figure 1: Taxonomy for the 3D structure of VS methods [25].

Decision tree (DT) is represented as an actual tree with its root at the top and the leaves at the bottom. The root of the tree is divided into two or more branches. The branch may be broken down into two branches or more. This process continues until the leaf is reached, meaning no more split remains.

Multilayer perceptron (MP) has two main types of artificial neural networks (ANNs), which are supervised and unsupervised networks. Every network consists of a series of linked neurons. A neuron takes multiple numerical inputs and outputs of values depending on the number of inputs weighted. Popular functions of transformation embody the functions of tanh and sigmoid. Neurons are formed into layers. ANN can contain several hidden layers, and the neurons will only be linked to those in the next layers, known as forwarding feed networks, multiplayer perceptions (MLPs), or functional radial base network (RBN) [29].

Logistic regression (LR) is one of the simplest and frequently utilized ML algorithms for two-class classification. It is simple to implement and can be utilized as the baseline for any binary classification problem. In deep learning, basic principles are also constructive. The relationship between a single dependent relative binary variable and independent variables is defined and estimated by logistic regression. It is also a mathematical method for predicting binary classes. The effect or target variable is dichotomous in nature. Dichotomous means that only two possible groups can be used for cancer detection issues, for instance [30].

Gradient boosting is a type of ML boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error in the space of possible predictions for each training case [30].

In [31], the authors compared four Weka classifiers, including SVM, J48 tree, NB, and RF. From a completed cost-sensitive survey, SVM and Tree C4.5 (J48) performed well, taking minority group sizes into account. It shows that a hybrid of majority class undersampling and SMOTE can improve overall classification performance in an imbalanced dataset. In addition, the authors in [32] have employed a repetitive SVM as a sample method that is used for SVM processing from bioassay information on luciferase inhibition that has a high active/inactive imbalance ratio (1/377). The models' highest performance was 86.60 and 88.89 percent for active compounds and inactive compounds, respectively, associated with a combined precision of 87.74 percent when using validation and blind test. These findings indicate the quality of the proposed approach for managing the intrinsic imbalance problem in HTS data used to cluster possible interference compounds to virtual screening utilizing luciferase-based HTS experiments.

Similarly, in [33], the authors analyzed several common strategies for modeling unbalanced data, including multiple undersampling, threshold, ratio 1 : 3 undersampling, one-sided undersampling, similarity undersampling, cluster undersampling, diversity undersampling, and only threshold selection. In total, seven methods were compared using HTS datasets extracted from PubChem. Their analysis led to the proposal of a new hybrid method, which includes both low-cost learning and less sampling approaches. The authors claim that the multisample and hybrid methods provide accurate predictions of results more than other methods. Besides, some other research studies used toxicity datasets in imbalance algorithms as in [34]. The model was based on a data ensemble, where each model sees an equal distribution of the two toxic and nontoxic classes. It considers various aspects of creating computational models to predict cell toxicity based on cell proliferation screening dataset. Such predictive models can be helpful in evaluating cell-based screening outcomes in general by bringing feature-based data into such datasets. It also could be utilized as a method to recognize and remove potentially undesirable compounds. The authors concluded that the issue of data imbalance hindered the accuracy of the critical activity classification. They also developed an artificial random forest group model that was designed to mitigate dataset misalignment in predicting cell toxicity.

The authors in [35] used a simple oversampling approach to build an SVM model classifying compounds based on the expected cytotoxic versus Jurkat cell line. Oversampling with the minority has been shown to contribute to better predictive SVM models in training groups and external test groups. Consequently, the authors in [36] analyzed and validated the importance of different sampling methods over the nonsampling method in order to achieve well-balanced sensitivity and specificity of the ML model that has been created for unbalanced chemical data. Additionally, the study conducted an accuracy of 93.00% under the curve (AUC) of 0.94, a sensitivity of 96.00%, and specificity of 91.00% using SMOTE sampling and random forest classification to predict drug-induced liver injury. Although it was presented in the literature that some of the proposed approaches have succeeded somehow in responding to the issues of unbalanced PubChem datasets, there is still a lack of time efficiency during calculations.

3. Proposed KSMOTE Framework

K-Mean Synthetic Minority Oversampling Technique (KSMOTE) is proposed in this paper as a solution for virtual screening to drug discovery problems. KSMOTE combines *K*-mean and SMOTE algorithms to avoid the imbalanced original datasets, ensuring that the number of minority samples is as close as possible to the majority of the population samples. As shown in Figure 2, the data were first separated into two sets—one set contains majority samples and the other set contains the entire minority sample. First, majority samples were clustered into *K* clusters and minority samples, where *K* is greater than one in both cases. The number of clusters for each class is chosen according to the elbow method. The Euclidean distance was employed to calculate the distance between the center of each majority cluster and the center of each minority cluster. Each majority cluster sample was combined with the minority cluster sample subset to make *K* separate training datasets. This combination was done based on the largest distance between each majority and minority cluster. SMOTE was then applied to each combination of the clusters. It generates an instance of synthetic minority, oversampling minority class. For any minority example, the *k* (5 in SMOTE) is the nearest. Neighbors of the same class are determined, and then some instances are randomly selected according to the oversampling factor. After that, new synthetic examples are generated along the line between the minority example and its nearest chosen example.

3.1. Environment Selection and the Dataset. Since we are dealing with big data, a big data framework must be chosen. One of the most powerful frameworks that have been used in many data analytics is Spark. Spark is a well-known cluster computing engine that is very reliable. It presents application programming interfaces in various programming languages such as Java, Python, Scala, and R. Spark supports in-memory computing, allowing handling records much faster than disk-based engines Hadoop. Spark engine is advanced for in-reminiscence processing as well as disk-based totally processing [37]. It has been installed on different operating systems such as Windows and Ubuntu.

This paper implements the proposed approach using PySpark version 2.4.3 [38] and Hadoop version 2.7, installed on Ubuntu 18.04, and Python is used as a programming language. A Jupyter notebook version 3.7 was used. The computer configuration for experiments was a local machine Intel Core i7 with 2.8 GHz speed and 8 GB of RAM. To illustrate the performance of the proposed framework, three datasets are chosen. They are carefully chosen where each of them differs in its imbalance ratio. They are also large enough to illustrate the big data analysis challenge. The three datasets are AID 440 [39], AID 624202 [40], and AID 651820 [41]. All of them are retrieved from the PubChem database Library [5]. The three datasets are summarized in Table 1 and briefly described in the following paragraphs. All of the data exist in an unstructured format as SDF files. Therefore, they require some preprocessing to be accepted as input to the proposed platform:

(1) AID 440 is a formylpeptide receptor (FPR). The G-protein, coupled with the formylpeptide receptor, was one of the originating chemo-attracting receptor members [39]. It consists of 185 active and 24,815 nonactive molecules.

(2) AID 624202 is a qHTS test to identify small molecular stimulants for BRCA1 expression. BRCA1 has been involved in a wide range of cellular activities, including repairing DNA damage, cell cycle checkpoint control, growth inhibition, programmed cell death, transcription regulation, chromatin recombination, protein presence, and autogenously stem cell regeneration and differentiation. The increase in BRCA1 expression would enable cellular differentiation and restore tumor inhibitor function, leading to delayed tumor growth and less aggressive and more treatable breast cancer. Promising stimulants for BRCA1 expression could be new preventive or curative factors against breast cancer [40].

(3) AID 651820 is a qHTS examination for hepatitis C virus (HCV) inhibitors [41]. About 200 million people globally are hepatitis C (HCV) contaminated. Many infected individuals progress to chronic liver disease, including cirrhosis, with the risk of developing hepatic cancer. There is no effective hepatitis C vaccine available to date. Current interferon-based therapy is effective only in about half of patients and is associated with significant adverse effects. It is estimated that the fraction of people with HCV who can complete treatment is no more than 10 percent. The recent development of direct-acting antivirals against HCV, such as protease and polymerase inhibitors, is promising. However, it still requires a combination of peginterferon and ribavirin for maximum efficacy. Moreover, these agents are associated with a high resistance rate, and many have significant side effects (Figure 3)

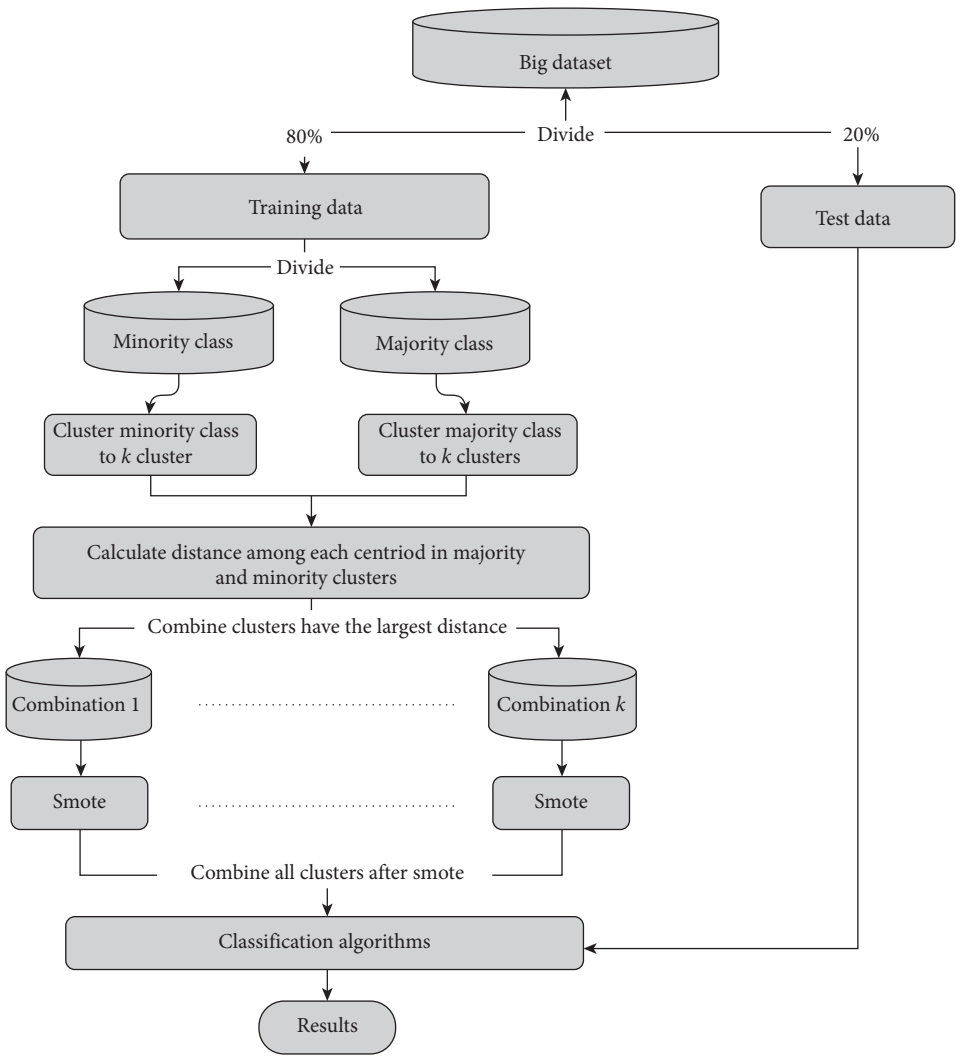


Figure 2: The proposed model.

Table 1: Benchmark imbalanced dataset.

Datasets	Total records	Inactive	Active	Balance ratio
AID 440	25,000	24,815	185	1 : 134
AID 624202	377,550	364,035	3980	1 : 91.5
AID 651820	283,005	271,341	11,664	1 : 23

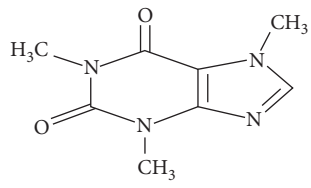


Figure 3: Molecules (chemical compound) [42].

3.2. Preprocessing. A chemical compound (molecules) is stored in SDF files, so those files have to be converted to feature vectors [$d1$, $d2$, and $d3$... $d1444$] with 1444 dimensions, and it was suitable to use PaDEL cheminformatics software [42] for this process. There are two types of PaDEL cheminformatics software, numeric and fingerprint descriptors. A PaDEL numeric descriptor gives information about the quantity of a feature in each compound. Molecules are represented based on constitutional, topological, and geometrical descriptors as well as other molecular properties. This includes aliphatic ring count, aromatic ring count, logP, donor count, polar surface area, and Balaban index. The Balaban index is a real number and can be either positive or negative. PaDEL is also used as a fingerprint descriptor, and it gives 881 attributes. The fingerprint was also calculated in order to compare the model performance derived from PaDEL descriptors. Molecules are referred to as instances and labeled as 1 or 0, where 1 means active and 0 means not active. The reader is directed to [43] for further information about the descriptors and fingerprints.

4. Evaluation Metrics

Instead of utilizing complicated metrics, four intuitive and functional metrics (specificity, sensitivity, G-mean, and accuracy) were introduced according to the following

reasons: First, the predictive power of the classification method for each sample, particularly the predictive power of the minority group (i.e., active power), is demonstrated by measuring performance for both sensitivity and specificity. Second, G-mean is a combination of sensitivity and specificity, indicating a compromise between the majority and minority output of the classification. Poor quality in predicting positive samples reduces the G-mean value, whereas negative samples are classified with accuracy with a high percentage. This is a typical state for imbalanced data collection. It is strongly recommended that external predictions be used to build a reliable model of prediction [41, 44]. The four statistical assessment methods are described as follows:

(1) Sensitivity: the proportion of positive samples appropriately classified and labeled, and it can be determined by the following equation:

$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{2}$$

where true positive (TP) corresponds to the right classification of positive samples (e.g., in this work, active compounds); true negative (TN) corresponds to the correct classification (i.e., inactive compounds) of negative samples; false positive (FP) means that negative samples have been incorrectly identified in positive samples; and false negative (FN) is an indicator to incorrectly classified positive samples.

(2) Specificity: the proportion of negative samples that are correctly classified; its value indicates how many cases that are predicted to be negative and that are truly negative as stated in the following equation:

$$\text{specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}, \tag{3}$$

(3) G-mean: it offers a simple way to measure the model's capability to correctly classify active and inactive compounds by combining sensitivity and specificity in a single metric. G-mean is a measure of balance accuracy [45] and is defined as follows:

$$\text{G-mean} = \sqrt{\text{specificity} \times \text{sensitivity}}. \tag{4}$$

G-mean is a hybrid of sensitivity and specificity, indicating a balance between majority and minority rating results. Low performance in forecasting positive samples also contributes to reducing G-mean value, even though negative samples are highly accurate. This is a typical unbalanced dataset condition [45].

(4) Accuracy: it shows the capability of a model to correctly predict the class labels as given in the following equation:

$$\text{accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{5}$$

4.1. Experimental Results. This section presents the results based on extensive sets of experiments. The average of the experiments is concluded and presented with discussion. Tenfold cross-validation is used to evaluate the performance of the proposed model. Besides, the original sample retrieved from the datasets is randomly divided into ten equal-sized subsamples. Out of the ten subsamples, one subsample is maintained as validation data for model testing, while the remaining nine subsamples are used as training data. The validation process is then replicated ten times (folds), using each of the ten subsamples as validation data exactly once. It is then possible to compute the average of the ten outcomes from the folds. To validate the effectiveness of KSMOTE, the performance of the proposed model is compared with that of SMOTE only and no-sampling models. Furthermore, two types of descriptors, numeric PaDEL and fingerprint, are used to validate their impact on the model's performance for all compilations. The results presented in this work are based on original test groups only without oversampling. Two types of descriptors, PaDEL and fingerprint, are used to generate five algorithms (RF, DT, MLP, LG, and GBT) to validate their effect on model output. Therefore, the performance of applying these algorithms is examined.

4.2. G-Mean-Based Performance. In this section, G-mean results are presented for the three selected datasets. Figure 4 describes G-mean for the AID 440 dataset based on both PaDEL and fingerprint descriptors. This figure demonstrates the performance of the various PaDEL descriptor and fingerprint sets. It also shows a comparison between three different approaches, which are no-sample, SMOTE, and KSMOTE. Besides the performance of employing RF, DT, MLP, LG, and GBT classifiers with the three different approaches examined, 20% of the datasets are used for testing, as mentioned before.

As shown in Figure 4, the best G-mean gained in the case of PaDEL fingerprint descriptor was by the KSMOTE, where it reaches, on average, 0.963. However, utilizing KSMOTE with LG gives almost G-means of 0.97, which is the best value over other classifiers. Based on our experiments, SMOTE- and no-sample-based PaDEL fingerprint descriptors are not recommended for virtual screening and classification where their G-mean is too small.

With the same settings applied to the fingerprint descriptor, the PaDEL numeric descriptor performance is examined. Again, the results are almost similar, where the KSMOTE shows higher performance than SMOTE and no-sample with nearly 55%. However, KSMOTE with DT and GBT classifiers are not up to other classifiers' level in this case. Therefore, it is recommended to utilize RF, MLP, and LG when a numeric descriptor is used. On the contrary, although the performance of SMOTE and no-sample is not that good, they show enhancement over the PaDEL fingerprint with an average of 10%. Among the overall results, it has been noticed that the worst G-mean value is produced from applying RF classifier with no-sample approach.

The performance of KSMOTE produced from the AID 440 dataset is confirmed using the AID 624202 dataset. As shown in Figure 5, KSMOTE gives the best G-mean using all

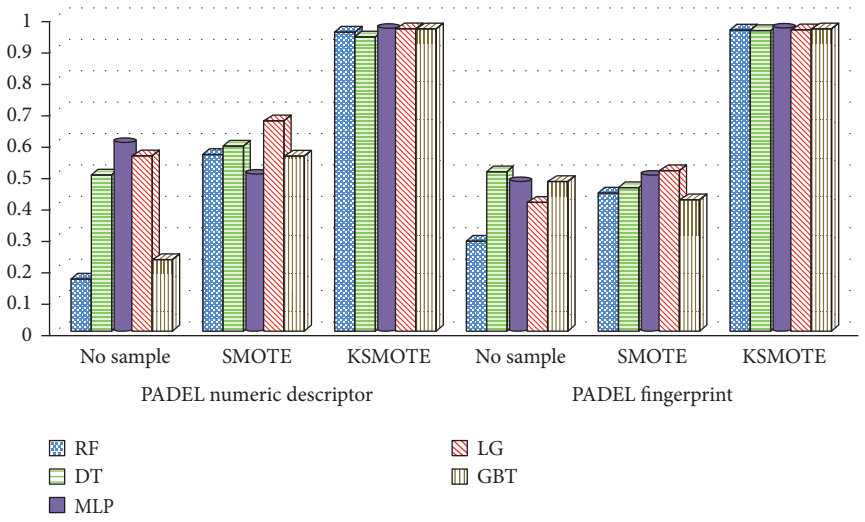


Figure 4: G-mean of PaDEL descriptor and fingerprint for AID 440.

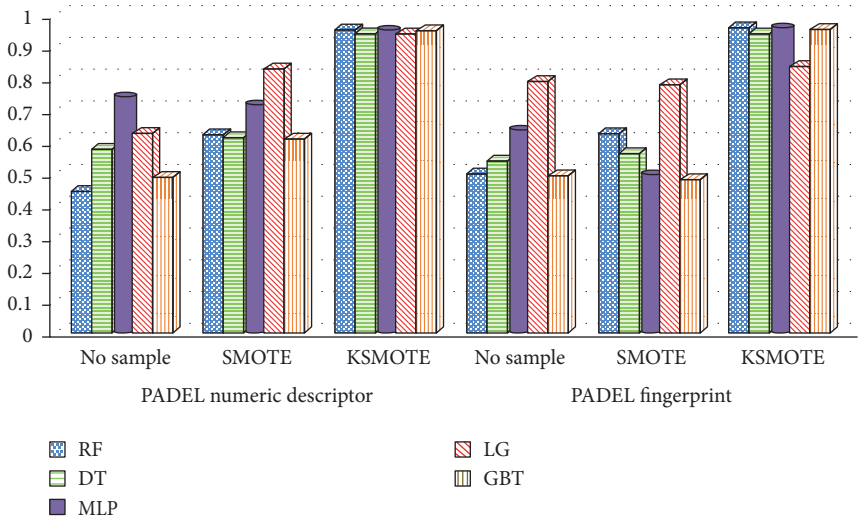


Figure 5: G-mean of PaDEL descriptor and fingerprint for AID624202.

of the classifiers. The only drawback shown is the G-mean of LG performance by almost 8% from other classifier results in case of PaDEL fingerprint classifier is used. On the contrary, LG classifier shows much more progress than before, where its average G-mean reached 0.8, which is a good achievement. For the rest of the classifiers in both descriptors, G-mean values are enhanced from the previous dataset. Still, among all the classifiers, G-mean value for the RF classifier with no-sample approach is worst. The overall conclusion is that KSMOTE is recommended to be used with both AID 440 and AID 624202 datasets.

Table 2 summarizes Figures 4–6 collecting all of the results in one place. It shows the complete set of experiments and average G-mean results in values to see the overall picture. Again, as can be exerted from the table, the proposed KSMOTE approach gives the best results of G-mean on the three datasets. It is believed that partitioning active and nonactive compounds to K clusters and then combining pairs that have large distances led to an accurate rate of oversampling instances in the SMOTE algorithm. This explains why the proposed model produces the best results.

4.3. Sensitivity-Based Performance. Sensitivity is another metric to measure the performance of the proposed approach compared to others. Here, the sensitivity performance presentation is a little bit different where the performance of the three approaches is displayed for the three datasets. Figure 7 shows the sensitivity of all datasets based on PaDEL numeric descriptor, while Figure 8 presents the sensitivity results for the fingerprint descriptor. It is obvious that the KSMOTE sensitivity values are superior to other approaches using both descriptors. In addition, SMOTE is overperforming the no-sample approach in almost all of the cases. For the AID 440 dataset, a low sensitivity value of 0.37 for the minority class is shown by the MLP model from the initial dataset (i.e., without SMOTE resample). The SMOTE algorithm was introduced to

Table 2: Complete set of experiments and average G-mean results.

Algorithm		PaDEL numeric descriptor				PaDEL fingerprint			
		No-sample	SMOTE	KSMOTE	Time	No-sample	SMOTE	KSMOTE	Time
AID 440	RF	0.167	0.565	0.954	23	0.29	0.442	0.96	12
	DT	0.5	0.59	0.937	**9.3**	**0.51**	0.459	0.958	4.9
	MLP	**0.6**	0.5	**0.963**	20	0.477	0.498	**0.964**	**9.6**
	LG	0.56	**0.67**	**0.963**	11	0.413	**0.512**	0.96	5.6
	GBT	0.23	0.56	**0.963**	33	0.477	0.421	0.963	17.1
AID624202	RF	0.445	0.625	**0.952**	29.7	0.5	0.628	**0.96**	**15.3**
	DT	0.576	0.614	0.94	10	0.54	0.564	0.94	5
	MLP	**0.74**	0.715	0.95	25.2	0.636	0.497	0.958	13.5
	LG	0.628	**0.83**	0.94	**26.8**	**0.791**	**0.78**	0.837	13.25
	GBT	0.489	0.61	0.95	45	0.495	0.482	0.954	22.36
AID 651820	RF	0.722	0.792	0.956	41	0.741	0.798	**0.92**	19.25
	DT	0.725	0.72	0.932	8.78	0.765	0.743	0.89	4.44
	MLP	**0.82**	0.817	0.915	**35**	**0.788**	**0.8**	0.91	17.3
	LG	0.779	**0.8357**	**0.962**	19	0.75	0.768	0.89	**9.36**
	GBT	0.714	0.742	0.9	60.5	0.762	0.766	0.905	29.9

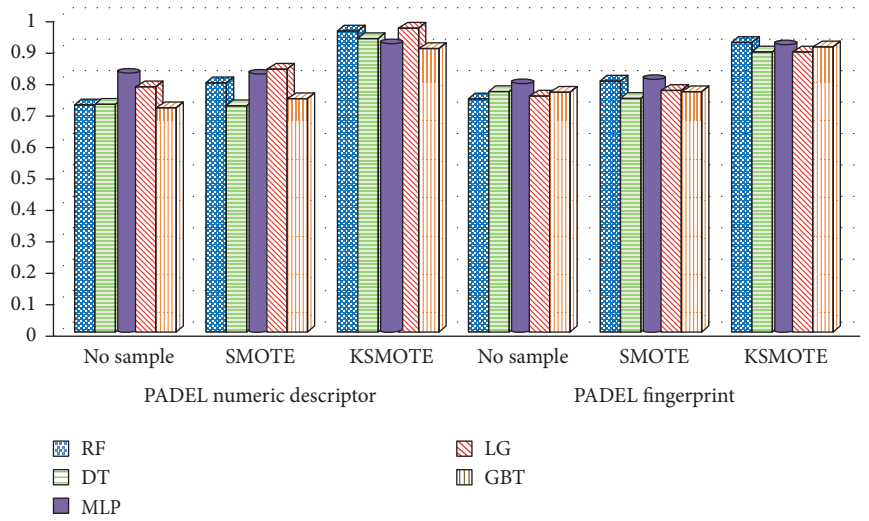


Figure 6: G-mean of numeric PaDEL descriptor and fingerprint for AID 65182.

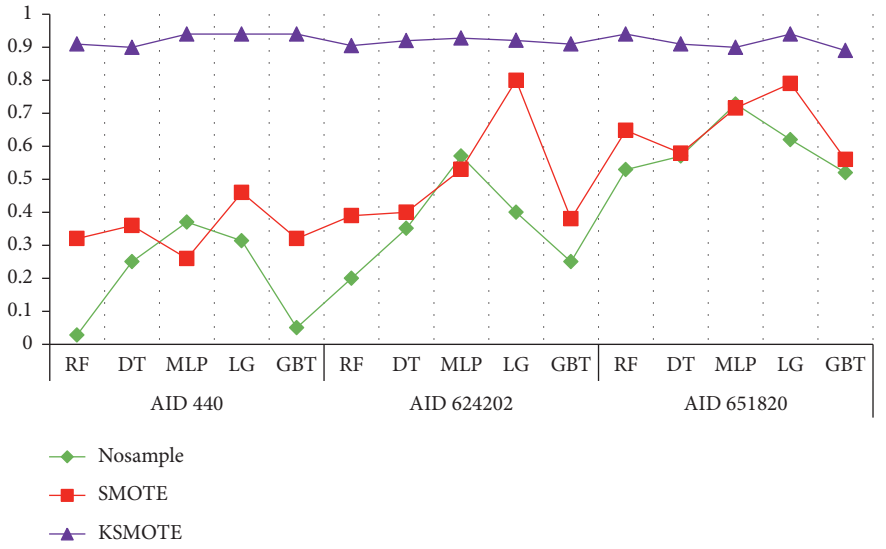


Figure 7: Sensitivity of all datasets for the PaDEL numeric descriptor.

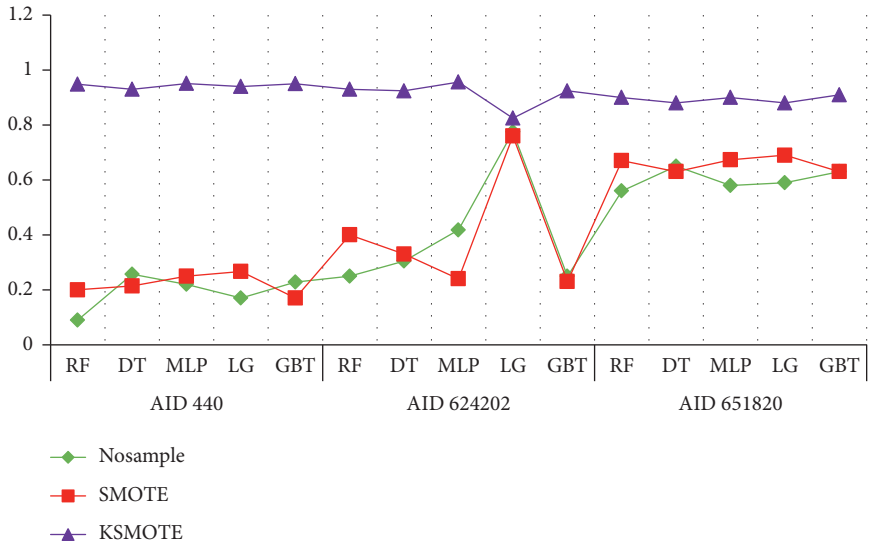


Figure 8: Sensitivity of all datasets for the fingerprint descriptor.

oversampling this minority class to significantly improve perceptibility, with LG jumping from 0.314 to 0.46%. The KSMOTE algorithm has been used to sample this minority class to boost the predictability of the interesting class, which represents active compounds. Besides, the sensitivity increases have been shown in the LG, with the KSMOTE sensitivity value jumped from 0.46 to 0.94 percent.

For the AID 624202 dataset, where the original model hardly recognizes the rare class (active compounds) with exceptionally weak sensitivity, a 0.2 RF is more significantly improved. However, the sensitivity value improves dramatically from 0.4 to 0.8 in LG with the incorporation of SMOTE. In KSMOTE, however, the sensitivity value in MLP rises considerably from 0.8 to 0.928. The model classification of AID 651820 is similarly enhanced, with sensitivity in the majority class in MLP 0.728 (inactive compounds).

Figure 8 presents the sensitivity of the three approaches using fingerprint descriptors. It is obvious that KSMOTE sensitivity values are superior to other approaches. As can be seen, the performance differs based on the type of the used dataset; on the contrary, KSMOTE has a stable performance using different classifiers. In other words, the difference in the KSMOTE is not that noticeable. However, it is clear from the figure that, on average, SMOTE and no-sample approaches have the same performance as well as behavior when applied to all datasets. Besides, the sensitivity results became much better when they are applied on the AID651820 dataset than when AID 440 and AID624202 were used. Again, the results go along with the previous measurement.

4.4. Specificity-Based Performance. Specificity is another important performance measure where it measures the percentage of negative classified classes that are correctly classified. Figures 9 and 10 show the specificity of all classifiers using PaDEL and fingerprint descriptors, respectively. Those figures illuminate two points as follows:

(A) All algorithms, on average, are correctly identifying the negative classes, except SMOTE, LG classifier in both AID 624202 and AID 651820 datasets

(B) Fingerprint descriptor results are more stable than PaDEL descriptor results

To summarize the sensitivity and specificity results, Table 3 shows the produced results using different classifiers. KSMOTE has a superior result in most of the experiments. Sensitivity and specificity results of the three datasets in numeric and fingerprint descriptors are shown, and the values marked in bold are the highest gained values among the results. Those values show the efficiency of the proposed method, KSMOTE.

4.5. Computational Time Comparison. One of the issues that the algorithms always face is the computational time, especially if those algorithms are designed to work on limited-resource devices. The models without SMOTE, for samples with minority classes, cannot achieve adequate performance. On the other hand, the five classifiers' computational time when KSMOTE is used has been proven to be accurate using sensitivity and G-mean values in almost all of the three PubChem datasets (see Figures 4–8). The computed computational time is reported in Figures 11 and 12. It is interesting to note that both PaDEL descriptors and fingerprint PaDEL descriptors produce similar computational efficiency among the five classifiers. From the results, DT and LG give the best computation time among all classifiers, followed by MLP. The computational time values of fingerprint PaDEL are much smaller than the numeric PaDEL descriptor's computational time in most cases. However, looking at the maximum computational time among the classifiers in both Figures 11 and 12, it turns out to be a GBT classifier with a value of 29 and 60 seconds in numeric and fingerprint descriptors.

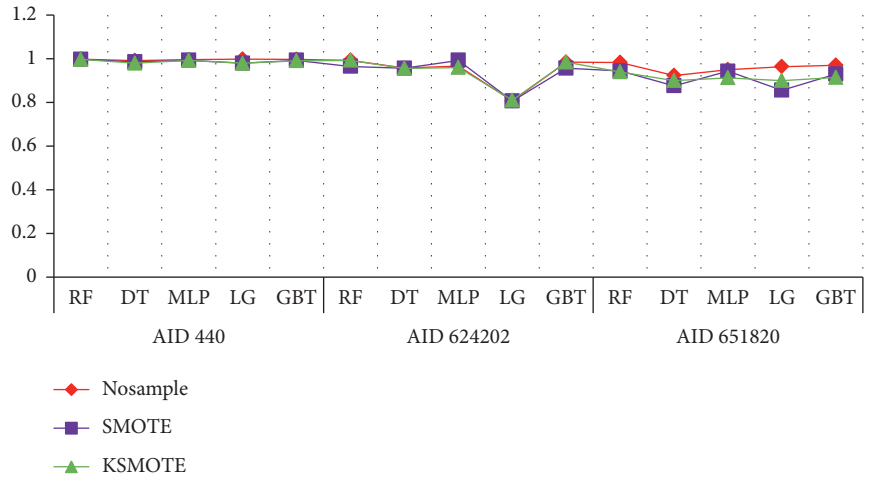


Figure 9: Specificity of all datasets for the PaDEL descriptor.

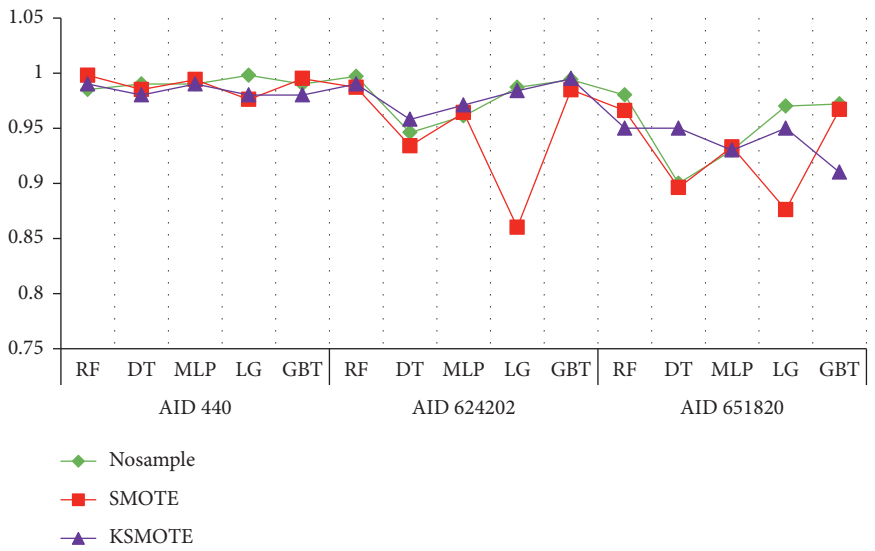


Figure 10: Specificity of all datasets for the fingerprint descriptor.

5. Discussion

It has been considered that the key problem of HTS data is its extreme imbalance, with only a few hits, often identified from a wide variety of compounds analyzed. The imbalance ratio distribution of AID 440 is 1/134, that of AID 624202 is 1/91.5, and for AID 651820, it is 1/23, as shown in Figure 1. Based on the conducted experiments, our proposed model successfully distinguished the active compounds with an average accuracy of 97% and the inactive compounds with an accuracy of 98%, with a G-mean of 97.5%.

Moreover, HTS data size, which typically comprises hundreds of thousands of compounds, poses another challenge. A statistical model may be trained and optimized on such a highly time-intensive dataset. Big data platforms, such as Spark in this study, were computationally effective and dramatically decreased computing costs in the optimized phase and substantially improved the KSMOTE model's performance.

Ideally, the KSMOTE model separates active (minority) dataset from inactive (majority) data with maximum distance. But the KSMOTE model, constructed from an imbalanced dataset, appears to move the hyperplane away from the optimal location to the minority side. Thus, most items are likely to be categorized into the majority class by both no-sample and SMOTE models, leading to a broad difference between specificity and sensitivity. Therefore, such a model's predictability can be significantly weak. We not only rely on cluster sampling to investigate the progress of the KSMOTE model but also built a SMOTE model for each sampling round.

We checked the KSMOTE model's performance with the blind dataset, which included 37 active compounds and 4963 inactive compounds for AID 440. KSMOTE in AID 440 was able to classify the inactive compounds very well with an overall accuracy of >98 percent, while it correctly classifies the active compounds at an accuracy of 95%. However, AID 624202 contains 796 active compounds and 72807 inactive

Table 3: Sensitivity and specificity results of the three datasets in numeric and fingerprint descriptors.

Algorithm		PaDEL numeric descriptor						PaDEL fingerprint					
		No-sample		SMOTE		KSMOTE		No-sample		SMOTE		KSMOTE	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
AID 440	RF	0.028	**0.998**	0.32	**0.997**	0.91	**0.997**	0.09	0.985	0.2	0.998	0.948	**0.99**
	DT	0.25	0.992	0.36	0.986	0.9	0.98	**0.257**	0.99	0.214	0.985	0.93	0.98
	MLP	**0.37**	0.996	0.26	0.993	0.94	0.993	0.22	0.99	0.25	0.994	**0.951**	**0.99**
	LG	0.314	**0.998**	**0.46**	0.98	0.94	0.98	0.17	**0.998**	**0.267**	0.976	0.94	0.98
	GBT	0.05	0.997	0.32	0.993	0.94	0.991	0.228	0.99	0.17	**0.995**	0.95	0.98
AID 624202	RF	0.2	**0.993**	0.39	0.965	0.905	**0.993**	0.25	0.997	0.4	**0.987**	0.93	0.99
	DT	0.351	0.958	0.4	0.957	0.92	0.956	0.305	0.946	0.33	0.934	0.924	0.958
	MLP	**0.57**	0.967	0.53	**0.993**	**0.928**	0.96	0.418	0.961	0.24	0.964	**0.956**	0.971
	LG	0.4	0.807	**0.8**	0.806	0.921	0.81	**0.775**	0.987	**0.76**	0.86	0.825	0.984
	GBT	0.25	0.985	0.38	0.957	0.91	0.985	0.249	**0.994**	0.23	0.9848	0.924	**0.995**
AID 651820	RF	0.529	0.983	0.648	**0.944**	**0.94**	**0.94**	0.56	**0.98**	0.67	0.966	0.9	**0.95**
	DT	0.57	0.923	0.579	0.876	0.91	0.9	**0.65**	0.9	0.63	0.896	0.88	**0.95**
	MLP	**0.728**	0.95	0.716	0.943	0.9	0.913	0.58	0.93	0.673	0.933	0.9	0.93
	LG	0.62	0.964	**0.79**	0.856	**0.94**	0.9	0.59	0.97	**0.69**	0.876	0.88	**0.95**
	GBT	0.52	**0.97**	0.56	0.93	0.89	0.914	0.63	0.972	0.63	**0.967**	**0.91**	0.91

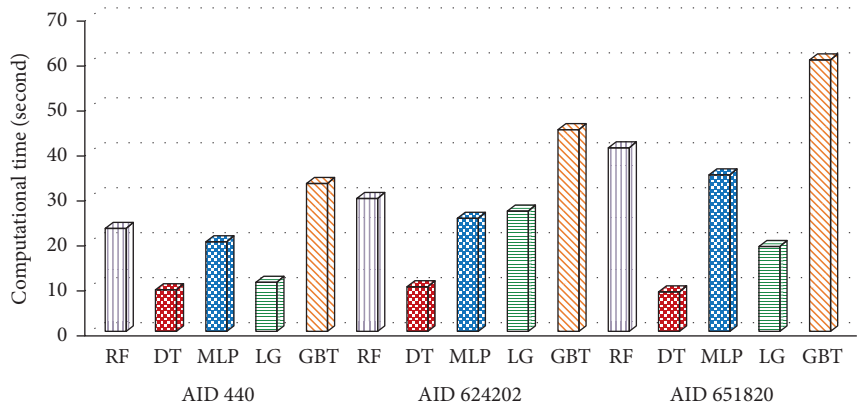


Figure 11: Comparison of computational time (seconds) in KSMOTE for the numeric PaDEL descriptor.

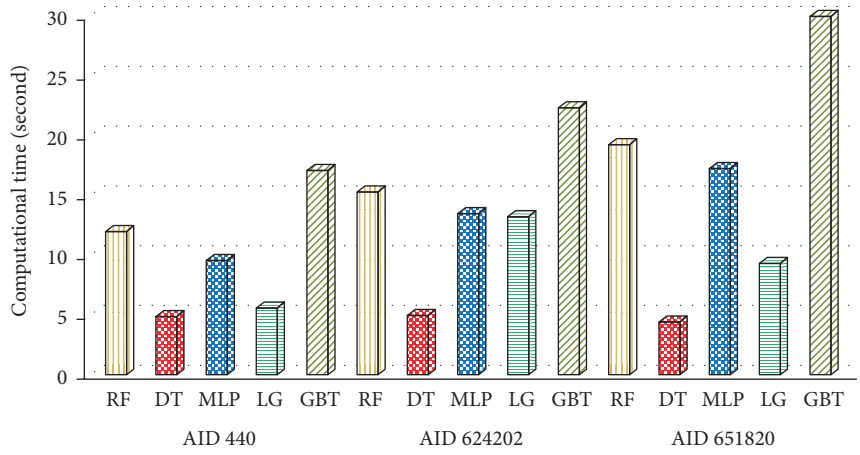


Figure 12: Comparison of computational time (seconds) in KSMOTE for the fingerprint PaDEL descriptor.

compounds, and AID 651820 contains 2332 active compounds and 54269 inactive compounds. Thus, AID 624202 is able to identify inactive compounds very well with an average accuracy of >96 percent, while it correctly classifies the active compounds at an accuracy of 94%.

Also, AID 651820 is able to identify inactive compounds quite well with an average accuracy of >94 percent, while it correctly classifies the active compounds at an accuracy of 92%. KSMOTE is considered better than the other systems (SMOTE only and no-sample) because it depends at the beginning on the lack of similarity in taking the samples of clusters. This dissimilarity between samples increases classifiers' accuracy, and besides that, it used the SMOTE to increase the number of minority samples (active compound) by generating a new sample for the minority.

The strength of KSMOTE lies in the fact that, in addition to the oversampling minority class accurately, CBOS produced new samples that do not affect majority class space in any way. We use the randomness in an effective way by restraining the maximum and minimum values of the newly generated samples.

Recent developments in technology allow for high-throughput scanning facilities, large-scale hubs, and individual laboratories that produce massive amounts of data at an unprecedented speed. The need for extensive information management and analysis attracts increasing attention from researchers and government funding agencies. Hence, computational approaches that aid in the efficient processing and extraction of large data are highly valuable and necessary. Comparison among the five classifiers (RF, DT, MLP, LG, and GBT) showed that DT and LG not only performed better but also had higher computational efficiency. Detecting active compounds by KSMOTE makes it a promising tool for data mining applications to investigate biological problems, mostly when a large volume of imbalanced datasets is generated. Apache Spark improved the proposed model and increased its efficiency. It also enabled the system to be more rapid in data processing compared to traditional models.

6. Conclusion

Building accurate classifiers from a large imbalanced dataset is a difficult task. Prior research in the literature focused on increasing overall prediction accuracy; however, this strategy leads to a bias towards the majority category. Given a certain prediction task for unbalanced data, one of the relevant questions to ask is what kind of sampling method should be used? Although various sampling methods are available to address the data imbalance problem, no single sampling

method works best for all problems. The choice of data sampling methods depends, to a large extent, on the nature of the dataset and the primary learning objective. The results indicate that, regardless of the datasets used, sampling approaches substantially affect the gap between the sensitivity and the specificity of the model trained in the nonsampling method. This study demonstrates the effectiveness of three different models for balanced binary chemical datasets. This work implements both *K*-mean and SMOTE on Apache Spark to classify unbalanced datasets from PubChem bioassay. To test the generalized application of KSMOTE, both PaDEL and fingerprint descriptors were used to construct classification models. An analysis of the results indicated that both sensitivity and G-mean showed a significant improvement after KSMOTE was employed. Minority group samples (active compounds) were successfully identified, and pathological prediction accuracy was achieved. In addition, models created with PaDEL descriptors showed better performance. The proposed model achieved high sensitivity and G-mean, up to 99% and 98.3%, respectively.

For future research, the following points are identified based on the work described in this paper:

(1) It is necessary to find solutions to other similar problems in chemical datasets, such as using semi-supervised methods to increase labeled chemical datasets. There is no doubt that the topic needs to be studied in depth because of its importance and its relationship with other areas of knowledge, such as biomedicine and big data.

(2) It is suggested to study deep learning algorithms for the treatment of class imbalance. Utilizing deep learning may increase the accuracy of the classification overcoming the deficiencies of existing methods.

(3) One more open area is the development of an online tool that can be used to try different methods and decide on the best results instead of working with only one method at a time.

Data Availability

The dataset used is publicly available at (1) AID 440 (https://pubchem.ncbi.nlm.nih.gov/bioassay/440), (2) AID 624202 (https://pubchem.ncbi.nlm.nih.gov/bioassay/624202), and (3) AID 651820 (https://pubchem.ncbi.nlm.nih.gov/bioassay/651820).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Kim and M. Shin, "An integrative model of multi-organ drug-induced toxicity prediction using gene-expression data," *BMC Bioinformatics*, vol. 15, no. S16, p. S2, 2014.

[2] N. Nagamine, T. Shirakawa, Y. Minato et al., "Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening," *PLoS Computational Biology*, vol. 5, no. 6, Article ID e1000397, 2009.

[3] P. R. Graves and T. Haystead, "Molecular biologist's guide to proteomics," *Microbiology and Molecular Biology Reviews*, vol. 66, no. 1, pp. 39–63, 2002.

[4] B. Chen, R. F. Harrison, G. Papadatos et al., "Evaluation of machine-learning methods for ligand-based virtual screening," *Journal of Computer-Aided Molecular Design*, vol. 21, no. 1-3, pp. 53–62, 2007.

[5] NIH, "PubChem," 2020, https://pubchem.ncbi.nlm.nih.gov/.

[6] R. Barandela, J. S. Sánchez, V. Garćia, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.

[7] L. Han, Y. Wang, and S. H. Bryant, "Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem," *BMC Bioinformatics*, vol. 9, no. 1, p. 401, 2008.

[8] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant, "PubChem as a public resource for drug discovery," *Drug Discovery Today*, vol. 15, no. 23-24, pp. 1052–1057, 2010.

[9] L. Cao and F. E. H. Tay, "Financial forecasting using support vector machines," *Neural Computing & Applications*, vol. 10, no. 2, pp. 184–192, 2001.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[11] C.-Y. Chang, M.-T. Hsu, E. X. Esposito, and Y. J. Tseng, "Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods," *Journal of Chemical Information and Modeling*, vol. 53, no. 4, pp. 958–971, 2013.

[12] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study1," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[13] G. M. Weiss and F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[14] K. Verma and S. J. Rahman, "Determination of minimum lethal time of commonly used mosquito larvicides," *The Journal of Communicable Diseases*, vol. 16, no. 2, pp. 162–164, 1984.

[15] S. Q. Ye, *Big Data Analysis for Bioinformatics and Biomedical Discoveries*, Chapman and Hall/CRC, Boca Raton, FL, USA, 2015.

[16] S. del Río, V. López, J. M. Benítez, and F. Herrera, "On the use of mapreduce for imbalanced big data using random forest," *Information Sciences*, vol. 285, pp. 112–137, 2014.

[17] A. Fernández, M. J. del Jesus, and F. Herrera, "On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9805–9812, 2009.

[18] K. Sid and M. Batouche, *Ensemble Learning for Large Scale Virtual Screening on Apache Spark*, Springer, Cham, Switzerland, 2018.

[19] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6585–6608, 2012.

[20] B. Hemmateenejad, K. Javadnia, and M. Elyasi, "Quantitative structure-retention relationship for the kovats retention indices of a large set of terpenes: a combined data splitting-feature

selection strategy," *Analytica Chimica Acta*, vol. 592, no. 1, pp. 72–81, 2007.

[21] S. Jain, E. Kotsampasakou, and G. F. Ecker, "Comparing the performance of meta-classifiers-a case study on selected imbalanced data sets relevant for prediction of liver toxicity," *Journal of Computer-Aided Molecular Design*, vol. 32, no. 5, pp. 583–590, 2018.

[22] A. V. Zakharov, M. L. Peach, M. Sitzmann, and M. C. Nicklaus, "QSAR modeling of imbalanced high-throughput screening data in PubChem," *Journal of Chemical Information and Modeling*, vol. 54, no. 3, pp. 705–712, 2014.

[23] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1–20, 2010.

[24] Y. Xiong, Y. Qiao, D. Kihara, H.-Y. Zhang, X. Zhu, and D.-Q. Wei, "Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates," *Current Drug Metabolism*, vol. 20, no. 3, pp. 229–235, 2019.

[25] B. K. Shoichet, "Virtual screening of chemical libraries," *Nature*, vol. 432, no. 7019, pp. 862–865, 2004.

[26] L. T. Afolabi, F. Saeed, H. Hashim, and O. O. Petinrin, "Ensemble learning method for the prediction of new bio-active molecules," *PLoS One*, vol. 13, no. 1, Article ID e0189538, 2018.

[27] A. C. Schierz, "Virtual screening of bioassay data," *Journal of Cheminformatics*, vol. 1, no. 1, p. 21, 2009.

[28] S. K. Hussin, Y. M. Omar, S. M. Abdelmageid, and M. I. Marie, "Traditional machine learning and big data analytics in virtual screening: a comparative study," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 47, pp. 72–88, 2020.

[29] l. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[31] Q. Li, Y. Wang, and S. H. Bryant, "A novel method for mining highly imbalanced high-throughput screening data in PubChem," *Bioinformatics*, vol. 25, no. 24, pp. 3310–3316, 2009.

[32] R. Guha and S. C. Schürer, "Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays," *Journal of Computer-Aided Molecular Design*, vol. 22, no. 6-7, pp. 367–384, 2008.

[33] P. Banerjee, F. O. Dehnbostel, and R. Preissner, "Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets," *Frontiers in Chemistry*, vol. 6, 2018.

[34] P. W. Novianti, V. L. Jong, K. C. B. Roes, and M. J. C. Eijkemans, "Factors affecting the accuracy of a class prediction model in gene expression data," *BMC Bioinformatics*, vol. 16, no. 1, p. 199, 2015.

[35] Spark, "Apache spark is a unified analytics engine for big data processing," 2020.

[36] P. AID440, "Primary HTS assay for formylpeptide receptor (FPR) ligands and primary HTS counter-screen assay for formylpeptide-like-1 (FPRL1) ligands," 2007, https://pubchem.ncbi.nlm.nih.gov/bioassay/440.

[37] P. 624202 AID, "qHTS assay to identify small molecule activators of BRCA1 expression," 2012, https://pubchem.ncbi.nlm.nih.gov/bioassay/624202.

[38] P. 651820 AID, "qHTS assay for inhibitors of hepatitis C virus (HCV)," 2012, https://pubchem.ncbi.nlm.nih.gov/bioassay/651820.

[39] Chem.libretexts.org, "Molecules and molecular compounds," 2020, https://chem.libretexts.org/Bookshelves/General_Chemistry/Map%3A_Chemistry_-_The_Central_Science_(Brown_et_al.)/02._Atoms_Molecules_and_Ions/2.6%3A_Molecules_and_Molecular_Compounds.

[40] X. Wang, X. Liu, S. Matwin, and N. Japkowicz, "Applying instance-weighted support vector machines to class imbalanced datasets," in *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data)*, pp. 112–118, IEEE, Washington, DC, USA, December 2014.

[41] V. H. Masand, N. N. E. El-Sayed, M. U. Bambole, V. R. Patil, and S. D. Thakur, "Multiple quantitative structure-activity relationships (QSARs) analysis for orally active trypanocidal N-myristoyltransferase inhibitors," *Journal of Molecular Structure*, vol. 1175, pp. 481–487, 2019.

[42] S. W. Purnami and R. K. Trapsilasiwi, "SMOTE-least square support vector machine for classification of multiclass imbalanced data," in *Proceedings of the 9th International Conference on Machine Learning and Computing - ICMLC*, pp. 107–111, Singapore, Singapore, Febuary 2017.

[43] T. Cheng, Q. Li, Y. Wang, and S. H. Bryant, "Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection," *Journal of Chemical Information and Modeling*, vol. 51, no. 2, pp. 229–236, 2011.

[44] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced datasets Knowledge and Information Systems," *Knowledge and Information Systems*, vol. 25, no. 1, p. 25, 2010.

[45] S. P. R. Trapsilasiwi, "SMOTE-least square support vector machine for classification of multiclass imbalanced data," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, pp. 107–111, ACM, Singapore, February 2017.

Hindawi
Complexity
Volume 2021, Article ID 6631837, 11 pages
https://doi.org/10.1155/2021/6631837

Research Article

A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records

Qiuli Qin,[1] **Shuang Zhao**,[1] **and Chunmei Liu**[2]

[1]*School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China*
[2]*Beijing Jiaotong University Health Service Center, Beijing 100044, China*

Correspondence should be addressed to Qiuli Qin; qlqin@bjtu.edu.cn

Received 10 December 2020; Revised 4 January 2021; Accepted 9 January 2021; Published 28 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

Because of difficulty processing the electronic medical record data of patients with cerebrovascular disease, there is little mature recognition technology capable of identifying the named entity of cerebrovascular disease. Excellent research results have been achieved in the field of named entity recognition (NER), but there are several problems in the pre processing of Chinese named entities that have multiple meanings, of which neglecting the combination of contextual information is one. Therefore, to extract five categories of key entity information for diseases, symptoms, body parts, medical examinations, and treatment in electronic medical records, this paper proposes the use of a BERT-BiGRU-CRF named entity recognition method, which is applied to the field of cerebrovascular diseases. The BERT layer first converts the electronic medical record text into a low-dimensional vector, then uses this vector as the input to the BiGRU layer to capture contextual features, and finally uses conditional random fields (CRFs) to capture the dependency between adjacent tags. The experimental results show that the F1 score of the model reaches 90.38%.

1. Introduction

Named entity recognition is to extract entities with actual meaning from massive unstructured text data [1, 2]. In the medical field, medical entities mainly include symptoms, examinations, diseases, drugs, treatments, operations, body parts etc., and are an important part of the establishment of a medical knowledge base. Chinese electronic medical record (CMR) [3] is a combination of structured and unstructured texts, which generally include not only patient information, but also a large amount of medical knowledge, but it is difficult to process. With the development of deep learning technology, entity recognition algorithms have applied research in many fields, but they lack applications in the field of cerebrovascular diseases (CVD) [4].

Cerebrovascular diseases have become one of the most threatening diseases to human health in the world due to the four characteristics [5, 6]. The treatment of cerebrovascular diseases highly depends on the doctor's experience. With the increase in the number of patients with CVD, there is a greater demand for cerebrovascular disease physicians. Since the training cycle of professional doctors is relatively long [7, 8], it will cause an imbalance in supply and demand of "more patients and fewer doctors". With the introduction of the concept of "AI + Medical," the use of machine learning technology to assist diagnosis and treatment, that is, through the construction of a complex model, the feedback mechanism is used to continuously optimize the parameters of the model then, use the existing clinical data and neuroimaging data in the hospital to diagnose and treat cerebrovascular diseases or predict recurrence. On the one hand, auxiliary diagnosis and treatment decision-making is helpful to improve the professional level of doctors and improve the quality of CVD medical services. On the other hand, it can optimize the uneven distribution of medical resources [9]. At present, the scientific research of machine learning in the field of CVD mainly focuses on two aspects: diagnosis and prognosis prediction of cerebrovascular disease: (1) From the perspective of CVD diagnosis, most scholars use structured data to nest machine learning models to complete

disease diagnosis. The literature [10–12] established a joint diagnosis model based on logitic regression method and XGBoost machine learning method by collecting clinical data of demographic characteristics. (2) From the perspective of prognosis prediction, the use of machine learning methods for risk prediction has gradually become the trend of disease prediction, while machine learning methods such as random forest, decision tree, SVM, and other machine learning methods have achieved certain research results in the prediction of cerebrovascular diseases. Literature [13–15] constructed logistic regression, k-NN, random forest, decision tree, and SVM machine learning models based on follow-up data, and verified the advantages of machine learning models in cerebrovascular disease risk prediction. It shows that the effect of neural network learning score is better. In short, from the analysis of clinical data sources, CVD medical data includes cerebrovascular disease imaging data, follow-up data, electronic medical records, and other data. The focus of most scholars is still on structured data such as follow-up data and neuroimaging data, while the focus on electronic medical record data in the field of CVD is slightly lower. At present, the increase in the number of CVD patients is accompanied by the ever-increasing number of electronic medical records for CVD patients. Electronic medical records can provide scholars with more data resources. For the processing of unstructured text information in electronic medical records, named entity recognition (NER) is a key step, and there are relatively few researches dedicated to named entity recognition in the field of cerebrovascular.

The current research on named entity recognition focuses on three aspects: (1) From the perspective of traditional entity recognition methods, traditional methods include methods based on dictionaries and rules [16–23]. This method relies heavily on domain dictionaries and domain experts. The selection of features is done manually, and subjectivity and labor costs are relatively large. With the development of machine learning technology [19, 20], more and more scholars are paying attention to models such as conditional random field (CRF), Hidden Markov Model (HMM), Support Vector Machine (SVM), However, NER based on traditional machine learning technology has higher requirements for feature selection [21–23], and the quality of feature selection directly affects the effect of entity recognition. (2) From the perspective of deep learning methods: with the development of deep learning technology, literature [24–27] confirmed the advantages of deep neural network technology by comparing traditional CRF models, that is, deep neural network technology has less artificial feature intervention than traditional methods and can obtain higher accuracy and recall rate. Deep learning can automatically extract word features, reduce the subjectivity of feature selection, and help further improve the accuracy of recognition results. Therefore, it is better than traditional statistical algorithms such as CRF and HMM. The common single-entity recognition neural network generally only considers the sample input and lacks in-depth thinking about the output relationship. Based on the idea of model fusion, most scholars usually use LSTM-CRF as the main framework to solve the deficiencies of neural network models. The available literature [28–32] uses traditional word2vec, Glove, and other word vector methods, uses a BiLSTM-CRF model as the core, and adds a CNN model, attention mechanism, RNN model, etc., to the core framework. Furthermore, the word vector undergoes continuous fine-tuning of the parameters, resulting in the final recognition achieving more accurate recognition. The process of parameter tuning is employed to set the hyperparameters of the model. However, for the BiLSTM-CRF model, there are more parameter settings, and the model training time is longer. Literature [26–31, 33, 34] proposed the BGRU neural model, which has simple results and high computational efficiency, can make full use of context information to eliminate entity ambiguity, and has some good effects in the field of entity recognition. (3) From the perspective of pre-training models, the above-described preprocessing models all use traditional word vector methods such as word2vec and Glove. This method focuses on the feature extraction between words and often ignores word context information. In order to improve this problem, as the Google BERT pre-training model is proposed, the literature studies [35–38] combine the BERT word embedding model on the basis of the traditional BiLSTM-CRF model and consider the polysemy of a word in combination with the context. The P value, R value, and F1 score have all been improved. It can be seen that BERT has strong semantic analysis capabilities.

In order to solve the problems of ignoring context information, low model efficiency, and susceptibility to word segmentation in electronic CVD medical entity recognition processing. We propose a BERT-BiGRU-CRF neural network model to identify named entities in electronic medical records of cerebrovascular diseases. Specifically, the BERT layer first converts the electronic medical record text into a low-dimensional vector, then inputs the vector into the BiGRU layer to capture contextual features, and finally uses a CRF to capture dependency between adjacent tags. The entity extraction model proposed in this paper has achieved good recognition results.

2. BERT-BiGRU-CRF Model Construction

In the NER field, the use of deep neural network models for entity recognition has become the mainstream. This article uses BiGRU-CRF as a benchmark to extract named entities in the field of cerebrovascular diseases. The reason why the BERT pre-training language model is chosen is that the text vector is used as the input of the model, and the granularity of Chinese division is divided into character-level and word-level. Existing research shows that character-level pretraining schemes show better results [37, 39], while the BERT pre-training language model is a character-level pretraining program. That is, each word in the text is converted into a vector by querying the word vector table, as the model input; the model output is the vector representation combined with the context.

The overall structure of the BERT-BiGRU-CRF model is shown in Figure 1. The model is mainly divided into 3 layers. The first layer is the BERT layer. Through the BERT pre-training language model, each word in the sentence is converted into a low-dimensional vector form. The second layer is the BiGRU layer, which aims to automatically extract semantic and temporal features from the context. The third layer is the CRF layer, which aims to solve the dependency between the output tags to obtain the global optimal annotation sequence of the text.

In this study, the named entity recognition model was used to identify the medical named entity in the electronic medical record of cerebrovascular disease. The specific steps are as follows:

(1) EMR data preprocessing, that is, processing the original electronic medical record text data set and express the electronic medical record text set as $H = \{h_1, h_2, \ldots, h_n\}$, where the i-th electronic medical record text is expressed as $h_i = < w_{i1}, w_{i2}, \ldots, w_{in} >$. Predefined entity category $C = \{c_1, c_2, \ldots, c_m\}$ is divided and annotated according to character-level, and the characters and predefined categories are separated by spaces when annotated.

(2) Construct the electronic medical record text training data set.

(3) Model training, that is, training the BERT-BiGRU-CRF named entity recognition model. Take the electronic medical record test text collection $D_{\text{test}} = \{d_1, d_2, \ldots, d_N\}$ as input, and take the entity and its corresponding category pair as output: $\{\langle m_1, c_1 >, < m_2, c_2 >, \ldots, < m_p, c_p \rangle\}$, where the entity $mi = < h_i, b_i, e_i >$ represents the entity that appears in the document; $h_i, b_i, and\ e_i$, respectively, represent the start and end positions of m_i in h_i, and no overlap between entities is required, ie $e_i < b_i + 1$. C_{mi} represents the predefined category of the entity m_i, then calculates the $F1$ score according to the precision rate and the recall rate, and uses the $F1$ score as the model comprehensive evaluation index.

2.1. BERT Pre-training Language Model. Bidirectional Encoder Representation from Transformers (BERT) [40] is an unsupervised and deep bidirectional language representation model for pre-training. In order to accurately represent the context-related semantic information in the EMR, it is necessary to call the interface of the model to obtain the embedded representation of each word in the electronic medical record. BERT uses the deep two-way transformer encoder as the main structure of the model. Transformer introduces the self-attention mechanism and also draws on the residual mechanism of the convolutional neural network, so the training speed of the model is fast and the expression ability is strong. And also abandoning the RNN loop structure, the overall structure of the BERT model is shown in Figure 2.

En is the coded representation of the word, Trm is the transformer structure, and Tn is the word vector of the target word after training. The operating principle of the model is to use the transformer structure to construct a multi-layer bidirectional Encoder network, which can read the entire text sequence at one time, so that each layer can integrate the contextual information. The input of the BERT model adopts the embedding addition method. By adding three vectors, Token Embeddings, Segment Embeddings, and Position Embeddings, the purpose of pre-training and predicting the next sentence is achieved. In Chinese electronic medical record text processing, the semantics of characters or words in different positions have different semantics. Transformer indicates that the information embedded in the sequence of the tag sequence is its relative position or absolute position information, as shown in the following formulae:

$$PE\left(P_{\text{pos}}, 2i+1\right) = \cos\left(\frac{P_{\text{pos}}}{1000^{(2i/d\text{model})}}\right), \tag{1}$$

$$PE\left(P_{\text{pos}}, 2i\right) = \sin\left(\frac{P_{\text{pos}}}{1000^{(2i/d\text{model})}}\right), \tag{2}$$

where P_{pos} is the position of the word in the text, i represents the dimension, and d_{model} is the dimension of the encoded vector. The odd position is encoded using the cosine function. Even positions are coded using a sine function.

In order to better capture word-level and sentence-level information, the BERT pre-training language model is jointly trained by two tasks: Masked Language Model and Next Sentence Prediction. The Masked LM model [36] is similar to cloze filling. 15% of the words in the random mask corpus are marked with the "MASK" form, and then the BERT model is used to correctly predict the masked words. The strategy adopted in the training is that for 15% of the words, only 80% of the words are actually replaced with [mask], 10% of the words will be randomly replaced with other words, and the remaining 10% are unchanged. The Next SP model is to train the model to understand the relationship between sentences, that is, to judge whether the next sentence is the next sentence of the previous sentence. The specific method is to randomly select 50% correct sentence pairs from the text corpus, and 50% randomly select sentence pairs to judge the correctness of the sentence pairs. The Masked LM word processing and Next SP sentence processing are jointly trained to ensure that the information is represented by the vector of each word, so the model is comprehensive and semantically accurate. It fully depicts the characteristics of the character-level, word-level, sentence-level and even the relationship between sentences and increases the generalization ability of the BERT model.

2.2. BiGRU Layer. Gated Recurrent Unit (GRU) [34] gated recurrent unit structure is a variant of long and short-term memory neural network (LSTM). The LSTM structure includes forget gates, input gates and output gates. In traditional recurrent neural network (RNN) training, gradient disappearance or explosion problems often occur. LSTM

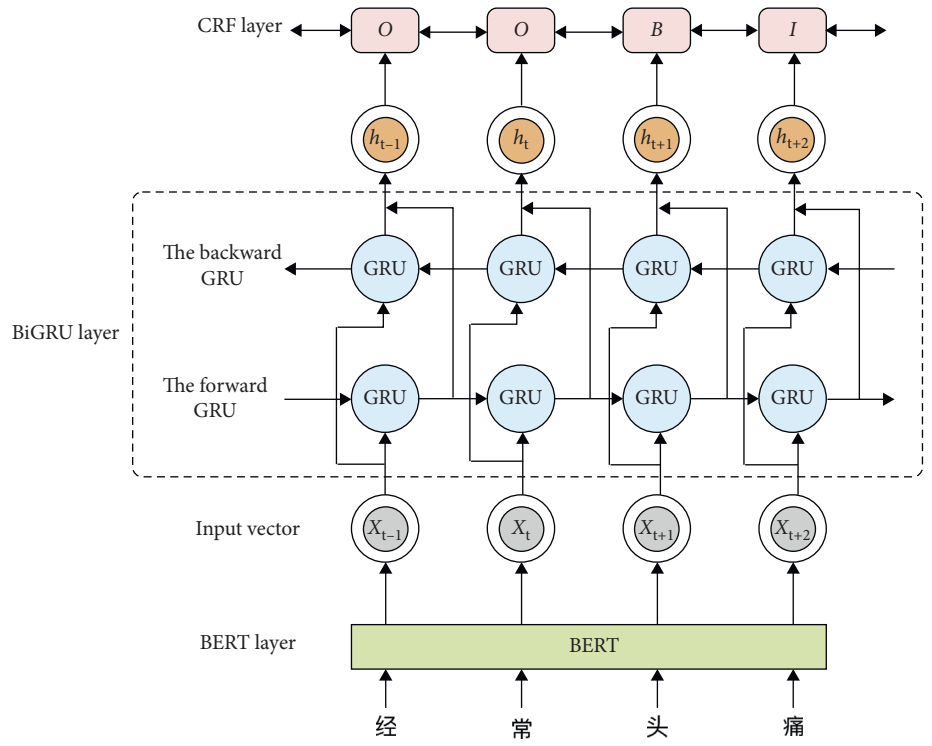


Figure 1: BERT-BiGRU-CRF model structure.

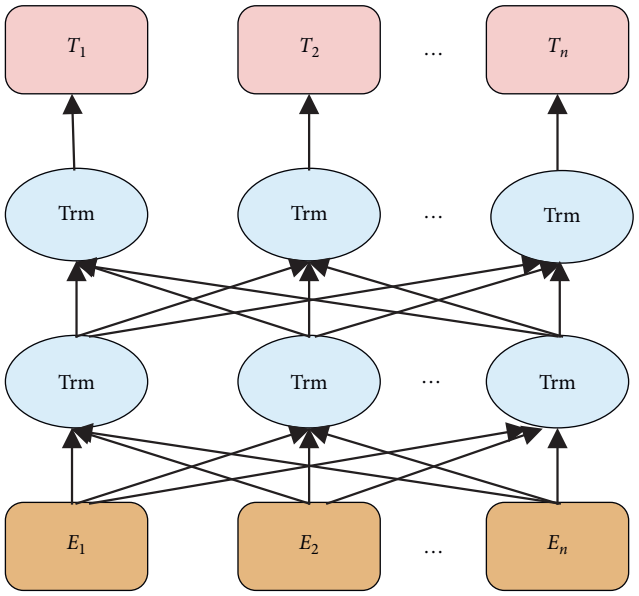


Figure 2: BERT network model.

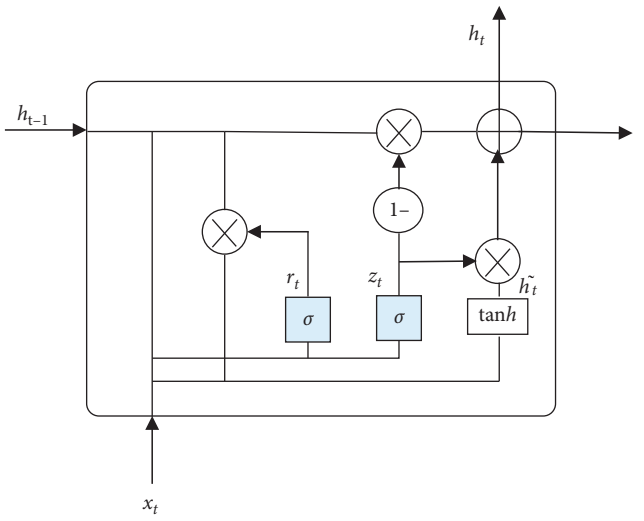


Figure 3: GRU structural unit.

only solves the problem of gradient disappearance to a certain extent, and the calculation is time-consuming. The GRU structure includes an update gate and a reset gate, and the GRU combines the forget gate and the input gate in the LSTM into an update gate. Therefore, GRU not only has the advantages of LSTM, but also simplifies its network structure. In the task of entity recognition of cerebrovascular disease electronic medical record, GRU can extract features effectively. Its network structure is shown as in Figure 3.

In the GRU structure, the update gate is z and the reset gate is r. The update gate z_t is to calculate how much electronic medical record information of the previous hidden layer state h_{t-1} needs to be transmitted to the current hidden state h_t. If z_t takes the value [0, 1], it needs to be transmitted when it is close to 1, and the information needs to be ignored when the value is close to 0. The reset gate r_t calculation formula is similar to the update gate principle, but the weight matrix is different. The calculation of z_t and r_t is shown in formulae (3) and (4). First, the electronic medical record data x_t input at time t, the state h_{t-1} of the hidden layer at the previous time, and the corresponding weights

are, respectively, multiplied and added to the σ function. After the calculation of z_t and r_t is completed, the content that needs to be memorized at time t can be calculated. Secondly, use the reset gate to determine the hidden state of the electronic medical record at $t-1$. The information that needs to be ignored at time t. Then, input r_t, h_{t-1}, x_t, and use tanh function to calculate the candidate hidden strong state. Finally, $h_t$transfers the cerebrovascular disease electronic medical record information retained in the current unit to the next unit; that is, at time t, the product of z_t and $\widetilde{h}$ represents the cerebrovascular disease information that the hidden unit h_t needs to retain. The product of $(1-z_t)$ and h_{t-1} indicates how much information needs to be forgotten. The calculation is shown in formulae (5) and (6) for details:

$$z_t = \sigma\left(w_z * \left[h_{t-1}, x_t\right]\right), \tag{3}$$

$$r_t = \sigma\left(w_r * \left[h_{t-1}, x_t\right]\right), \tag{4}$$

$$\widetilde{h} = \tanh\left(w_h * r_t\, \mu_t\, h_{t-1}, x_t\right), \tag{5}$$

$$h_t = \left(1 - z_t\right)\mu h_t - 1 + z_t \mu \widetilde{h}_t, \tag{6}$$

where x_t is the input of the electronic medical record of cerebrovascular disease at time t and h_{t-1} is the state of the hidden layer at the previous time; h_t is the hidden state at time t; w is the weight matrix; w_z is the update gate weight matrix and w_r is the reset gate weight matrix; σ is the sigmoid nonlinear transformation function and tanh is the activation function; $\widetilde{h}$ is the hidden state of candidate.

From the operating principle of the GRU unit, it can discard some useless information, and the structure of the model is simple, which reduces the computational complexity. However, the simple GRU cannot fully utilize the context information of the electronic medical record. Therefore, this paper designs the backward GRU to learn the backward semantics, and the GRU neural network forwards and backwards to extract the key features of the named entity in the electronic medical record of cerebrovascular disease, namely, the BiGRU model. The specific structure is shown in Figure 4. Based on the GRU principle, forward GRU is to obtain the above semantic feature (h_t), and backward GRU is to obtain the following semantic features h_t, and finally, the above and the following semantic features are combined to get h_t. Refer to formulae (7) and (8) for details:

$$\overrightarrow{h}_t = G\overrightarrow{R}U\left(x_t\right), \tag{7}$$

$$\overleftarrow{h}_t = G\overleftarrow{R}U\left(x_t\right), \tag{8}$$

$$h_t = < \overrightarrow{h}_t, \overleftarrow{h}_t >, \tag{9}$$

where $\overrightarrow{h}_t$ is the hidden layer state, the purpose is to obtain the above information from the GRU; $\overleftarrow{h}_t$ is the hidden layer state, the purpose is to obtain the following information from the GRU; $G\overrightarrow{R}U\left(x_t\right)$ means that it is represented by features from front to back; $GRU\left(x_t\right)$ means that it is represented by the back-to-front feature; and the final hidden layer state of h_t is the feature of the electronic medical record report.

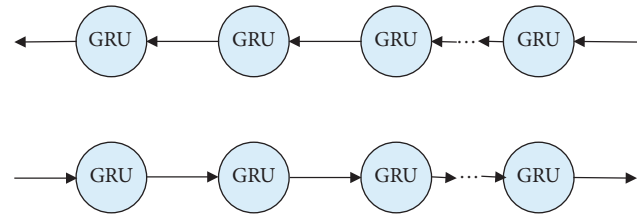


Figure 4: BiGRU structure.

2.3. CRF Layer. The NER problem can be regarded as a sequence labeling problem. The BiGRU layer outputs the hidden state context feature vector h, denoted as $h = (h_1, h_2, \ldots, h_t)$. This vector only considers the context information in the electronic medical record and does not consider the inter-label dependencies. Therefore, this paper adds a CRF layer to label the global optimal sequence and converts the hidden state sequence $h = (h_1, h_2, \ldots, h_t)$ into the optimal label sequence $y = (y_1, y_2, \ldots, y_t)$. CRF calculation principle [34]: firstly, for the specified electronic medical record input sequence $x = (x_1, x_2, \ldots, x_t)$, it calculates the score of each location, shown in formula (10). Secondly, calculate the probability of normalized sequence y through the Softmax function, shown in formula (11). Finally, the label sequence with the highest score is calculated using the Viterbi algorithm, shown in formula (12):

$$\text{score}\left(h, y\right) = \sum_{t=1}^{T} A_{yt-1}, yt + \sum_{t=1}^{T} W_{yt}^{t} h_t, \tag{10}$$

$$P\left(\frac{y}{h}\right) = \frac{e^{\text{score}\left(h,y\right)}}{\sum_{y' \in Y\left(h\right)} e^{\text{score}\left(h,y'\right)}}, \tag{11}$$

$$y* = \arg\max_{y' \in Y\left(h\right)} \left(h, y'\right), \tag{12}$$

where A is the transfer score matrix between tags; score (h, y) is the position score; W_{yt}^{T} is the parameter vector; $p\left(y/h\right)$ normalized probability function; $Y\left(h\right)$ represents all possible tag sequences; and formula (10) is to calculate the score (h, y) of each position in the input sequence from the output feature matrix of the BiGRU layer and the CRF transition matrix.

2.4. Training Process. The process of deep network model training is a process of repeatedly adjusting parameters so that loss reaches a minimum. However, due to the strong learning ability of deep network models, the problem of model generalization is prone to occur. For example, the problem of model under-fitting and over-fitting leads to poor adaptability of the model to new sample data. Therefore, regularization methods can generate many models with small parameter values. In other words, such

models have strong anti-interference ability and can adapt to different datasets and different "extreme conditions". It can increase the generalization capabilities of the model during the network training process. The method to solve the problem in this paper is the L2 regularization method, which can avoid the over-fitting problem, that is, adding regularization calculation to the cost function, shown in the following formula:

$$\text{loss} = E_{\text{in}} + \lambda \sum_{j} w_j^2, \tag{13}$$

where E_{in} is the training sample error that does not include the regularization term; λ is the adjustable parameter of regularization; and w_i represents the weight parameter.

3. Experimental Design

3.1. Data Preparation. The experimental data in this article was obtained from the Ai'ai medical electronic medical records website. The electronic medical record data is a total of 1,300 electronic medical records related to cerebrovascular diseases, which are composed of general patient information, chief complaint, medical history, physical examination, and diagnosis. In addition, this article sorts out the types of entities in the published papers related to named entities in electronic medical records, as shown in Table 1. According to the frequency of occurrence of entities in published literature, electronic medical record entities for cerebrovascular diseases are divided into five entities: disease, symptom, body part, examination, and treatment, which are also proposed by CCKS.

3.2. Data Preprocessing. The electronic medical record information is preprocessed, that is, line breaks and invalid characters, etc., are removed, and 36400 sentences are finally obtained. The testing dataset and the training dataset are divided into 2 : 8. The labeling system used in this article is BIO labeling. The five types of entities are disease, symptoms, body parts, examination, and treatment. Therefore, there are 11 labels, namely, O, Disease-B, Disease-I, Body-B, Body-I, Symptom-B, Symptom-I, Examination-B, Examination-I, Treatment-B, and Treatment-I. We conduct named entity labeling with doctors. Among the 300 medical records, we designated two annotators to annotate them at the same time, used Cohen's kappa to calculate the consistency of the annotations, and obtained a kappa value of 0.8. The labels to be predicted are shown in Table 2.

3.3. Experimental Settings. The experimental model in this article is built using tensorflow deep learning framework and Python programming language. The parameter update method is to update the parameters of the BiGRU-CRF model, and BERT is a fixed parameter. Table 3 lists the hyperparameter values of the experimental model in this article. These values have been modified according to relevant literature [34–38, 41] and have not been adjusted for the cerebrovascular electronic medical record data set in this article. The model parameter optimization in this paper adopts the stochastic gradient descent method (SGD), the initial learning rate is 0.015, the update of the learning rate adopts the step decay method, and the decay rate is 0.05. The model has achieved good experimental results in the training set and test set.

Table 1: Sorting of entities.

Entity category	Literature
Symptom	[28, 31, 32, 37, 39, 40, 42]
Treatment	[28, 31, 32, 37, 39, 40, 42]
Body part	[31, 32, 34, 42]
Examination	[28, 31, 32, 40]
Disease	[28, 31, 37, 40, 42]
Surgery	[34, 40]
Anatomy	[31, 32]
Independent symptoms	[31, 39]
Drug	[31, 39]

Table 2: BIO annotation example.

Text	Label	Meaning
头	Body-B	The beginning of the body part entity
部	Body-I	The middle of the body part entity
疼	Symptom-B	The beginning of the symptom entity
痛	Symptom-I	The middle of the symptom entity
3	O	Non-entity
小	O	Non-entity
时	O	Non-entity

Table 3: Parameter settings.

Parameter	Value
Dropout	0.5
Embedding	50
GRU-dim	200
Regularized intensity	1′10-8
Batch_size	64
Lr (learning rate)	0.015
LSTM_dim	200
Dr (decay rate)	0.05

3.4. Evaluation. This article uses the most commonly used evaluation index in the field of named entity recognition: precision rate (P), recall rate (R), and $F1$ score ($F1$). That is, P is the recognition rate of correctly recognized named entities, R is the rate of correctly recognized named entities in the test set, and $F1$ is the harmonic average of P and R, which is the comprehensive evaluation index of the model. Among them, the higher the P and R values, the higher the accuracy and the recall rate, but in fact, the two are contradictory in some cases. Therefore, the $F1$ score is often used to evaluate the overall performance of the model. The calculation formula is:

$$P = \frac{T_{\text{num\_entities}}}{S_{\text{num\_entities}}}, \tag{14}$$

$$R = \frac{T_{\text{num\_entities}}}{C_{\text{num\_entities}}}, \tag{15}$$

$$F1 = \frac{2P * R}{(P + R)}, \tag{16}$$

where $T_{\text{num\_entities}}$ is the number of correct entities identified; $S_{\text{num\_entities}}$ is the total number of all entities identified; and $C_{\text{num\_entities}}$ is the number of entities in the test set.

4. Experimental Results

4.1. Comparative Experiment Analysis. In order to prove the entity recognition effect based on the BERT-BiGRU-CRF model, this article uses BiLSTM-CRF as the baseline model, and the selected comparison models are BiLSTM-CRF, BERT-BiLSTM-CRF, and BERT-BiGRU-CRF. Model introduction:

(1) BiGRU-CRF model: this model inputs word vectors into the model for training.

(2) BiLSTM-CRF model: this model is a classic model in the NER field. It uses trained word vectors and then uses the BiLSTM-CRF model to extract entities.

(3) BERT-BiLSTM-CRF model: this model is based on the Google BERT model. Many scholars have embedded BERT in the BiLSTM-CRF model and achieved better recognition results in NER research.

4.2. Model Performance Comparison. The comparison model proposed in this paper first uses electronic medical record data for training and then uses a test set for testing. The specific comparison results are shown in Table 4.

From the comparison results of Table 4 and Figure 5, we can see that in terms of comprehensive evaluation indicators. In terms of precision rate, recall rate, and $F1$ score, the BERT-BiGRU-CRF model proposed in this article has increased by 2.9%, 5.0%, and 3.95%, respectively, compared with the BiGRU-CRF model. The difference between the two models is the embedding of BERT. It shows that BERT embedding can improve the recognition effect of entities. Compared with the BiLSTM-CRF model, the increase was 3.14%, 4.40%, and 4.34%, respectively. Compared with the BERT-BiLSTM-CRF model, the increase was 1.25%, 0.77%, and 1.01%, respectively, Therefore, all P, R, and $F1$ score are improved compared to the baseline model, indicating that the BERT-BiGRU-CRF model is more applicable to electronic medical record recognition in the CVD field. This is mainly due to the stronger ability of embedding BERT to extract features, which enables word vectors to fuse context information. On the other hand, the BiGRU-CRF model can input bidirectional information before and after the sequence, which can effectively avoid entity ambiguity.

Figure 6, in terms of entity types, horizontally compares the recognition effects of various entities under different models and compare BiLSTM-CRF, BERT-BiLSTM-CRF, and BiGRU-CRF. In terms of disease entities, they were increased by 9.87%, 2.73%, and 9.63%, respectively; on the symptom entity, they were increased by 1.62%, −0.33%, and 3.29%; on the body part entity, they were increased by 2.85%, 0.45%, and 3.10%. On the examination entity, they were increased by 0.76%, −0.25%, and 0.49%. In terms of the treatment entities, they were increased by 3.8%, 2.46%, and 3.29%, respectively. The overall recognition effect of different entities is compared longitudinally, the effect of checking entity recognition is higher than the comparison model, and the $F1$ score reaches 90%. However, the recognition effect of entities in the treatment category is relatively poor because the entities are relatively long and cannot clearly identify the boundaries of each entity. In short, the recognition effect of the BERT-BiGRU-CRF model proposed in this paper is higher than that of the control group.

4.3. Model Training Time. Model training is the process of parameter update. This article analyzes the relationship between the four models in the first 10 rounds of Epoch and $F1$. It can be seen from Figure 7 that the $F1$ score of the neural network model without BERT is continuously rising from a lower level, while the $F1$ score of the neural network model with BERT can be maintained at a higher level, and it takes iterations to reach the optimal $F1$ score, fewer times. In addition, as a whole, the $F1$ score of the BERT-BiGRU-CRF model is the highest. From the comparison of training time, Table 5 lists the time required for each model iteration. The training time of the BERT-BiGRU-CRF model for one round is 37 seconds shorter than that of the BERT-BiLSTM-CRF model. This is due to the simple structure of the BiGRU-CRF model and the higher efficiency of the model in calculation. In addition, comparing BiGRU-CRF and BERT-BiGRU-CRF models, it is worth noting that the BERT with full word mask is added to the neural network model, which improves the overall training efficiency of the model. The overall training efficiency is improved.

In summary, the BERT-BiGRU-CRF entity recognition model proposed in this paper has a better recognition effect than the control group. This model can make full use of context information, further avoid ambiguity, and effectively avoid repetition between entities, and the granularity of word segmentation in this article is small, which can improve the accuracy of entity recognition.

4.4. Entity Recognition Result. This paper uses the BERT-BiGRU-CRF named entity recognition model to identify 9393 entities (without deduplication). Among them, electronic medical records have the most descriptions of body parts, followed by symptoms and examination entities, while treatment and disease types are less. The specific results are shown in Figure 8.

Table 4: Comparison results of each model.

Model	Evaluation index (%)	Entity type					Comprehensive value
		Disease	Symptoms	Body parts	Examination	Treatment	
BiLSTM-CRF	*P*	84.01	89.09	85.79	91.16	83.28	86.67
	R	79.50	89.83	86.24	92.04	85.06	86.53
	*F*1	81.69	89.46	86.01	91.60	84.16	86.60
BERT-BiLSTM-CRF	*P*	88.64	90.38	87.82	91.98	84.12	88.59
	R	89.02	92.46	89.01	93.24	87.09	90.16
	*F*1	88.83	91.41	88.41	92.61	85.58	89.37
BiGRU-CRF	*P*	83.93	88.20	86.33	91.90	84.29	86.93
	R	80.02	87.38	85.20	91.84	85.22	85.93
	*F*1	81.93	87.79	85.76	91.87	84.75	86.43
BERT-BiGRU-CRF	*P*	90.65	90.26	88.64	92.33	87.29	89.83
	R	92.48	91.92	89.08	92.39	88.81	90.94
	*F*1	91.56	91.08	88.86	92.36	88.04	90.38

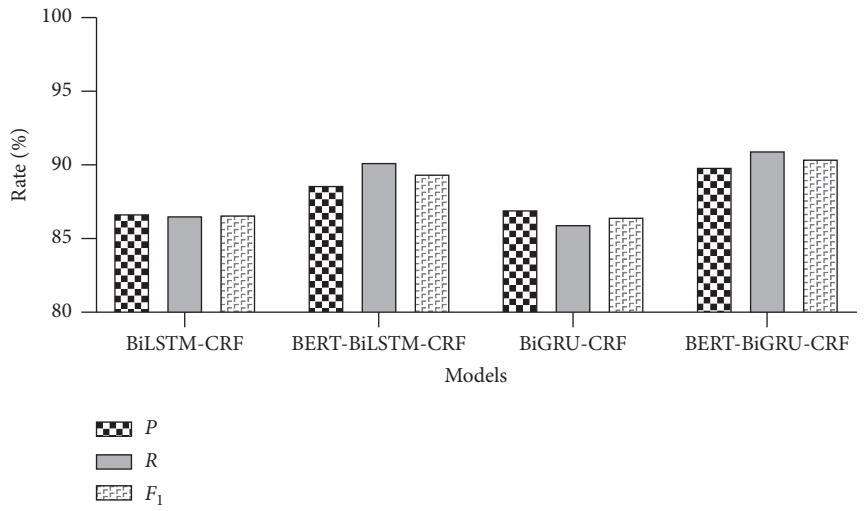


Figure 5: Comparison of various indicators of different models.

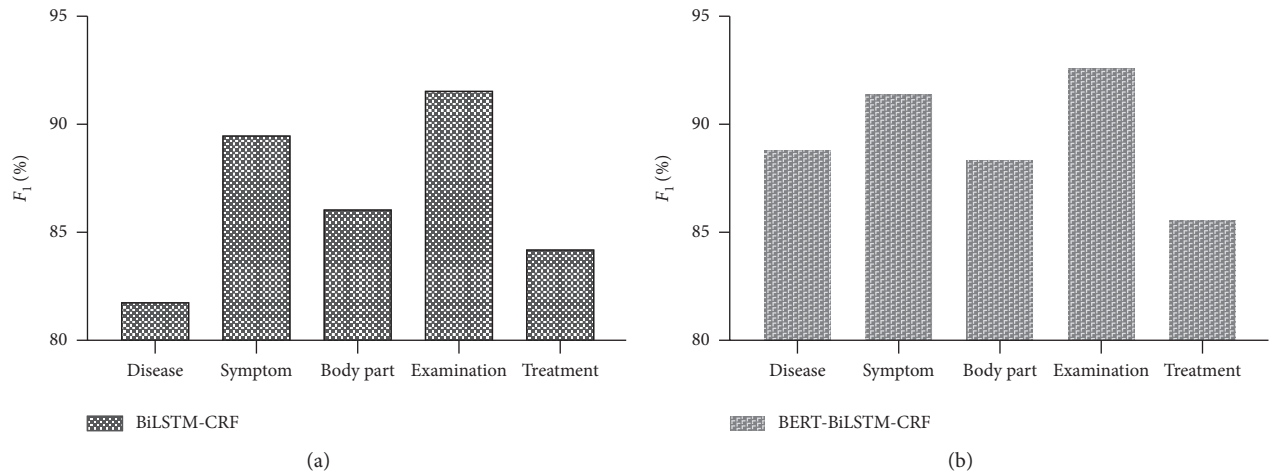


Figure 6: Continued.

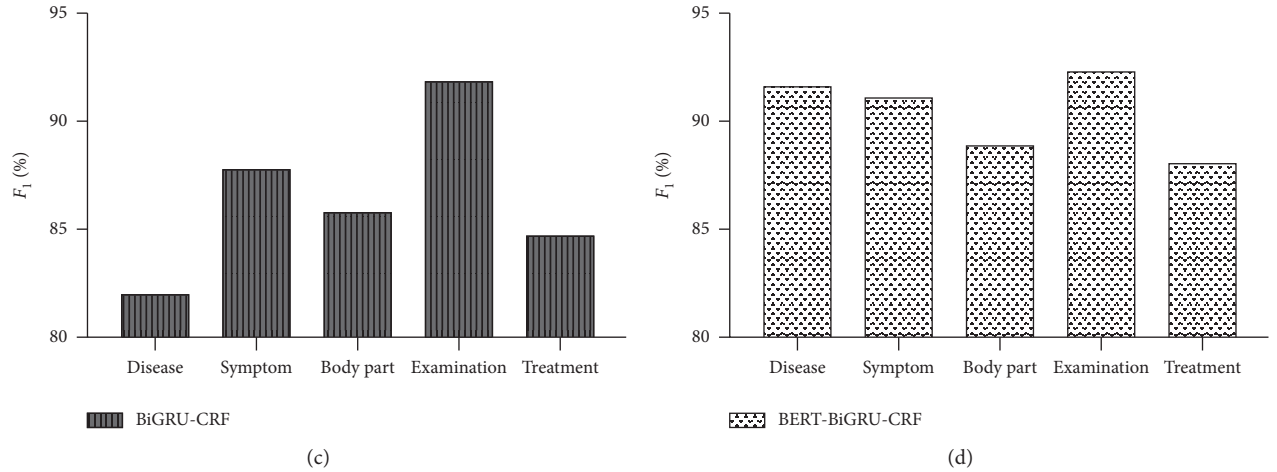


Figure 6: Comparison of entity $F1$.

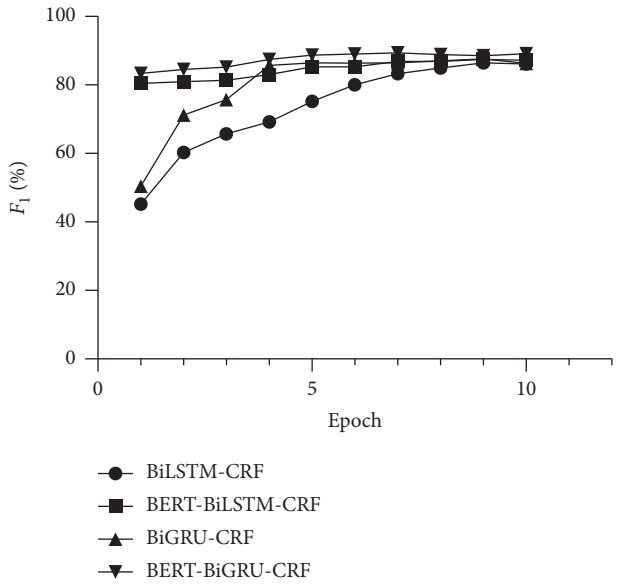


Figure 7: Epoch and $F1$ relationship.

Table 5: Time of one round of model training.

Model	Training time (s)
BiLSTM-CRF	472
BERT-BiLSTM-CRF	356
BiGRU-CRF	429
BERT-BiGRU-CRF	319

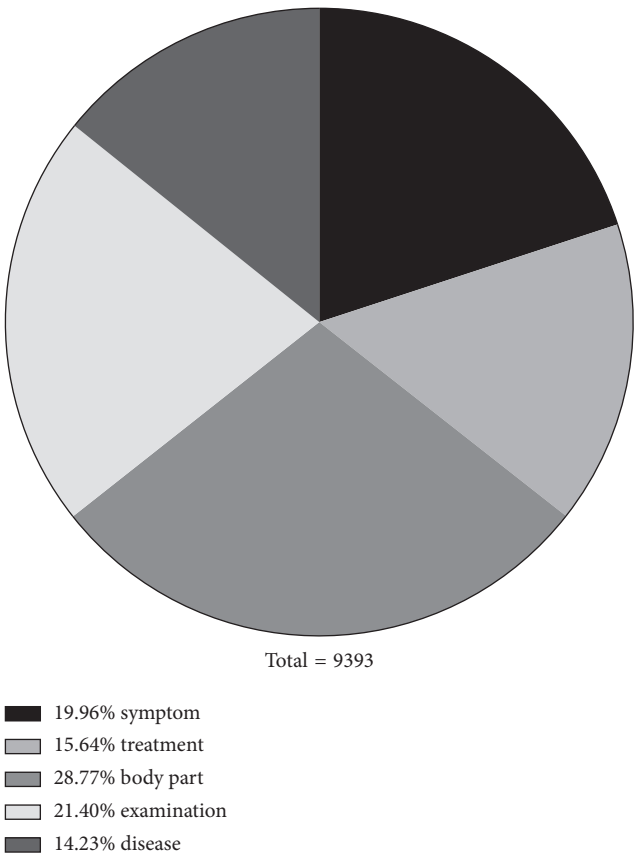


Figure 8: Distribution of various entities.

5. Conclusions

Aiming at the text data of electronic medical records of cerebrovascular diseases, this paper proposes a BERT-BiGRU-CRF entity recognition model to identify five key entities in the field of cerebrovascular diseases, which are "disease, symptoms, body parts, examination, and treatment." The model obtains the word vector combined with context information through the BERT layer and then obtains the optimal annotation sequence through the BiGRU-CRF neural network model. It not only guarantees a simple network structure and fast training speed but also can solve the problem of ambiguity in combination with context information. Next, on the one hand, we will study the construction of high-quality dictionaries. On the other hand, we will extract the relationship between different entities based on NER to construct a knowledge map in the field of cerebrovascular diseases, which is conducive to the further potential information of electronic medical records in the field of CVD.

Data Availability

The Chinese electronic medical record data used to support the findings of this study have been deposited in the Ai'ai medical repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Beijing Municipal Commission of Science and Technology Project (Z131100005613017), the model and demonstration application of the collaborative prevention and treatment of major diseases in Beijing's medical reform.

References

[1] H.-J. Jang and K.-O. Cho, "Applications of deep learning for the analysis of medical data," *Archives of Pharmacal Research*, vol. 42, no. 6, pp. 492–504, 2019.

[2] J. R. Ubbens and I. Stavness, "Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks," *Frontiers in Plant Science*, vol. 8, no. 1190, p. 2245, 2018.

[3] L. Shen, Q. Li, W. Wang et al., "Treatment patterns and direct medical costs of metastatic colorectal cancer patients: a retrospective study of electronic medical records from Urban China," *Journal of Medical Economics*, vol. 23, no. 5, pp. 456–463, 2020.

[4] Z. Li, T. Liu, L. Ding, Z. Liu, X. Li, and Y. Wang, "Research progress of machine learning in the diagnosis and treatment of cerebrovascular diseases," *Chinese Journal of Stroke*, vol. 15, no. 3, pp. 283–289, 2020.

[5] S. Wu, B. Wu, and M. Li, "Stroke in China: advances and challenges in epidemiology, prevention, and management," *Lancet Neurology*, vol. 18, no. 4, pp. 394–405, 2019.

[6] J. M. Wardlaw, C. Smith, and M. Dichgans, "Small vessel disease: mechanisms and clinical implications," *The Lancet Neurology*, vol. 18, no. 7, pp. 684–696, 2019.

[7] W. J. Powers, A. A. Rabinstein, T. Ackerson et al., "Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American heart association/American stroke association," *Stroke*, vol. 50, no. 12, pp. e344–e418, 2019.

[8] Z. Li, A. B. Singhal, and Y. Wang, "Stroke physician training in China," *Stroke*, vol. 48, no. 12, pp. e338–e340, 2017.

[9] E. Smit, G. Saposnik, G. Biessels et al., "Prevention of stroke in patients with silent cerebrovascular disease: a scientific statement for healthcare professionals from the American heart association/American stroke association," *Stroke*, vol. 48, no. 2, pp. E44–E71, 2017.

[10] J. Virtanen, M. Varpela, F. Biancari, J. Jalkanen, and H. Hakovirta, "Association between anatomical distribution of symptomatic peripheral artery disease and cerebrovascular disease," *Vascular*, vol. 28, no. 3, pp. 295–300, 2020.

[11] J. Virtanen, M. Varpela, and F. Biancari, "Association between anatomical distribution of symptomatic peripheral artery disease and cerebrovascular disease," *Vascular*, vol. 28, no. 32, pp. 95–300, 2020.

[12] A. Baluja, R. Moisés, A. Cordero et al., "Prediction of major adverse cardiac, cerebrovascular events in patients with diabetes after acute coronary syndrome," *Diabetes & Vascular Disease Research*, vol. 17, no. 1, pp. 67–69, 2019.

[13] C. Romana, T. Eva, D. Edin, J. Raka, and T. Milan, "Brain image segmentation based on firefly algorithm combined with K-means clustering," *Studies in Informatics and Control*, vol. 28, no. 2, pp. 167–176, 2019.

[14] B. Ambale-Venkatesh, X. Yang, C. Wu et al., "Cardiovascular event prediction by machine learning the multi-ethnic study

of atherosclerosis," *Circulation Research*, vol. 121, no. 9, p. 1092, 2017.

[15] I. Korvigo, M. Holmatov, A. Zaikovskii, and M. Skoblov, "Putting hands to rest: efficient deep cnn-rnn architecture for chemical named entity recognition with no hand-crafted rules," *Journal of Cheminformatics*, vol. 10, no. 28, 2018.

[16] H. Afify, K. Mohammed, and A. Hassanien, "Multi-images recognition of breast cancer histopathological via probabilistic neural network approach," *Journal of System and Management Sciences*, vol. 1, no. 2, pp. 53–68, 2020.

[17] A. Ekbal and S. Saha, "Simultaneous feature and parameter selection using multi-objective optimization: application to named entity recognition," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 4, pp. 597–611, 2020.

[18] Y. Meng and L. Long, "A deep learning Approach for a source code detection model using self-attention," *Complexity*, vol. 2020, Article ID 5027198, 15 pages, 2020.

[19] L. Gligic, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, "Named entity recognition in electronic health records using transfer learning bootstrapped neural networks," *Neural Networks*, vol. 121, pp. 132–139, 2020.

[20] C. Dumitrescu and I. Dumitrache, "Combining deep learning technologies with multi-level gabor features for facial recognition in biometric automated systems," *Studies in Informatics and Control*, vol. 28, no. 2, pp. 221–230, 2019.

[21] J. Chen, Z. Chang, and J. Xu, "Recognition and classification of biomedical named entities based on HMM," *Computer Age*, no. 10, pp. 40–42, 2006.

[22] M. Sui and L. Cui, "CRF model combining multiple characteristics for chemical substance-disease named entity recognition," *Modern Library and Information Technology*, no. 10, pp. 91–97, 2016.

[23] X. Sun, C. Sun, and F. Ren, "Biomedical named entity recognition based on deep conditional random fields," *Pattern Recognition and Artificial Intelligence*, vol. 29, no. 11, pp. 997–1008, 2016.

[24] A. Cocos, A. Fiks, and A. Masino, "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 813–821, 2017.

[25] F. Zhang and M. Wang, "Medical named entity recognition based on deep learning," *Computing Technology and Automation*, vol. 36, no. 1, pp. 123–127, 2017.

[26] L. Qin, N. Yu, and D. Zhao, "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 2, pp. 528–535, 2018.

[27] U. Zuperl and F. Cus, "A cyber-physical system for smart fixture monitoring via clamping simulation," *International Journal of Simulation Modelling*, vol. 18, no. 1, pp. 112–124, 2019.

[28] J. M. Giorgi and G. D. Bader, "Towards reliable named entity recognition in the biomedical domain," *Bioinformatics*, vol. 36, no. 1, pp. 280–286, 2020.

[29] I. Unanue, E. Borzeshi, and M. Piccardi, "Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition," *Journal of Biomedical Informatics*, vol. 76, pp. 102–109, 2017.

[30] L. Weston, V. Tshitoyan, J. Dagdelen et al., "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," *Journal of Chemical Information and Modeling*, vol. 59, no. 9, pp. 3692–3702, 2019.

[31] Z. H. Kilimci and S. Akyokus, "Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification," *Complexity*, vol. 2018, Article ID 7130146, 10 pages, 2018.

[32] S. Chowdhury, X. Dong, and L. Qian, "A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records," *BMC Bioinformatics*, vol. 19, no. 17, pp. 75–84, 2018.

[33] M. Al-Smadi, S. Al-Zboon, Y. Jararweh, and P. Juola, "Transfer learning for Arabic named entity recognition with deep neural networks," *Ieee Access*, vol. 8, no. 31, pp. 37–45, 2020.

[34] M. Gridach and H. Haddad, "Arabic named entity recognition: a bidirectional GRU-CRF approach," in *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, vol. 10761, pp. 264–275, , no. 1, Springer, Berlin, Germany, 2018.

[35] Y. Kim and T. Lee, "Korean clinical entity recognition from diagnosis text using bert," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 242, 2020.

[36] J. Lee, W. Yoon, S. Kim et al., "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics (Oxford, England)*, vol. 36, no. 4, pp. 1234–1240, 2020.

[37] P. Yang and W. Dong, "Chinese named entity recognition method based on BERT embedding," *Computer Engineering*, vol. 46, no. 4, pp. 40–45+52, 2020.

[38] Y. Cho and Y. Lee, "Biomedical named entity recognition using deep neural networks with contextual information," *BMC Bioinformatics*, vol. 20, no. 1, p. 735, 2019.

[39] X. Chen, C. Ouyang, Y. Liu, and Y. Bu, "Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, p. 2687, 2020.

[40] F. Li, Y. Jin, W. Liu, B. Rawat, P. Cai, and H. Yu, "Fine-tuning bidirectional encoder representations from transformers (Bert)-based models on large-scale electronic health record notes: an empirical study," *JMIR Medical Informatics*, vol. 7, no. 3, Article ID e14830, 2019.

[41] Y. Feng, H. Yu, and G. Sun, "Named entity recognition method based on BLSTM[J]," *Computer Science*, vol. 45, no. 2, pp. 261–268, 2018.

[42] J. Wu, Y. Cheng, and H. Hao, "Research on Chinese professional term extraction based on BERT embedded BiLSTM-CRF model," *Journal of Information*, vol. 39, no. 4, pp. 409–418, 2020.

Hindawi
Complexity
Volume 2021, Article ID 6649605, 10 pages
https://doi.org/10.1155/2021/6649605

Research Article

A Machine Learning Approach to Evaluate the Performance of Rural Bank

Jun Wei,[1] **Tao Ye**,[2] **and Zhe Zhang**[3]

[1]*School of Economics and Management, Beijing Jiaotong University, 100044 Beijing, China*
[2]*School of Finance, Capital University of Economics and Business, 100070 Beijing, China*
[3]*School of Management Science and Engineering, Shandong University of Finance and Economics, 250014 Jinan, China*

Correspondence should be addressed to Jun Wei; 15113140@bjtu.edu.cn

Received 9 December 2020; Revised 26 December 2020; Accepted 28 December 2020; Published 13 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

In the current performance evaluation works of commercial banks, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics. On the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction. This paper is the first to comprehensively investigate the predictability of multidimensional features on commercial bank performance using boosting regression tree. The dimensionality in the financial-related fields is relatively high. There are not only observable price data, financial fundamentals data, etc., but also many unobservable undisclosed data and undisclosed events; more sources of income cannot be explained by existing models. Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. In this method, a random subsample sampling is performed before training each regression tree. The adaptive reduction step size is used to replace the reduction step size setting of the original algorithm, which overcomes the shortcomings of low accuracy and poor generalization ability of the existing regression decision tree model. Compared to the BIRCH algorithm for classification of existing data, our proposed gradient boosting regression tree algorithm with adaptively reduced step size obtains better classification results. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance.

1. Introduction

The traditional Malmquist index [1] examines the efficiency and productivity changes of financial institutions. For example, Paradi et al. [2] estimated that the Bank of Canada will develop a two-stage DEA to simultaneously benchmark performance in different dimensions and modify the SBM.

Machine learning technology has certain applications in performance evaluation in the financial field. Taking the fund performance analysis and evaluation model as an example [3, 4], the use of related technologies can improve the traditional model evaluation methods required in the risk model and the fund performance evaluation method, such as the adjustment and optimization of the characteristic risks of individual stocks, the summary of potential laws, and the forecast adjustment of the fund's short-term exposure. In the current performance evaluation work of commercial banks using machine learning, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics; on the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction.

Most of the existing bank performance evaluation models are based on the Malmquist index method, but the information dimensionality in the financial-related fields is relatively high [5–8]. There are not only observable price data, financial fundamentals data, etc., but also many unobservable undisclosed data and undisclosed events; more sources of income cannot be explained by existing models.

Based on boosting regression tree technology, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. In this method, a random subsample sampling is performed before training each regression tree. The adaptive reduction step size is used to replace the reduction step size setting of the original algorithm, which overcomes the shortcomings of low accuracy and poor generalization ability of the existing regression decision tree model. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance. In this paper, we use predictive modeling and explanatory modeling in machine learning to evaluate the performance of rural banks and predict the possible development trend of performance. Explanatory models make assumptions about causality in advance and then use data to test them. Predictive models can unearth more complex laws in the data. However, the two are not completely opposed.

In summary, the contributions and innovations of this paper can be summarized as follows:

(1) This paper proposes the use of predictive models to evaluate the performance of commercial banks for the first time. In our opinion, compared with explanatory models, the predictive models can unearth more complex laws in the datasets.

(2) Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation.

(3) This study uses real commercial bank data from 30 provinces to conduct experiments. The experiment shows that the adaptively reduced step size gradient boosting regression tree algorithm proposed in this paper reveals the performance of commercial banks more objectively.

(4) This research not only uses predictive model methods to study bank performance evaluation from a more comprehensive perspective but also provides useful inspiration for commercial bank operations and management.

The rest of this paper is organized as follows: Section 2 is the related work part. Section 3 is the method proposed in this paper. Section 4 is the experimental results and analysis. We summarized a conclusion in Section 5.

2. Related Work

Ravi et al. presents a soft computing based bank performance prediction system [9]. It is an ensemble system whose constituent models are many kinds of neural network, SVM, CART, and a fuzzy rule-based classifier. Selecting a subset of favorable features is beneficial to improve the accuracy of bank performance prediction. For example, particle swarm optimization (PSO) is used to obtain suitable parameter settings for support vector machines (SVMs) and decision trees (DTs) [10], such as neural networks, support vector machines, and multicriteria decision aid that have also been used in the bank failure prediction, creditworthiness assessment, and underperformance [11].

A series of modeling techniques were employed to predict bank insolvencies on a sample of US-based financial institutions. The empirical results indicate that the method of random forests (RF) has a superior out-of-sample and out-of-time predictive performance, with neural networks also performing almost equally well as RF in out-of-time samples [12]. A sample of 3000 US banks (1438 failures and 1562 active banks) is investigated by two traditional statistical approaches (discriminant analysis and logistic regression) and three machine learning approaches (artificial neural network, support vector machines, and k-nearest neighbors) [13]. The empirical result reveals that the artificial neural network and k-nearest neighbor methods are the most accurate. An accurate risk assessment tool was proposed using unique KYC data and machine learning techniques to overcome problems in existing risk detection methods [14]. This work proposes that the bank branch is the best level at which to determine the degree of default risk and can also provide insight into patterns of suspicious transactions.

Several machine learning algorithms have been used on a real bank credit dataset for comparative analysis and to choose which algorithms are the best fit for learning bank credit data. These algorithms gave over 80% accuracy in prediction [15]. For evaluating bank efficiency and performance, a combined DEA with three machine learning approaches were used in 444 Ghanaian bank branches, decision-making units (DMUs). The results suggested that the decision tree (DT) and its C5.0 algorithm provided the best predictive model [16]. The potential usage of bagging also has been investigated which is one of the most popular ensemble learning methods, in building ensemble models, and is used to predict the determinants of Turkish IaDB profitability [17]. This empirical study indicates that bagging ensemble models are superior to their base learners and could improve the prediction accuracy of individual ML models.

There are a large number of empirical studies to analyse and evaluate machine learning techniques in the bank risk management [18]. The areas or problems in risk management also have been inadequately explored for further research. These prior empirical studies have shown that the application of machine learning in the management of banking risks such as credit risk, market risk, operational risk, and liquidity risk has been explored. For example, the combination of financial indicators, readability, sentiment categories, and bag-of-words was used to increase prediction accuracy. It shows that the quality of the prediction significantly increased when using the correlation-based feature selection of bag-of-words [19]. The supervised artificial neural network algorithm is implemented for classification purpose in customer retention and fraud detection [20].

We can clearly conclude that machine learning algorithms have been widely used in various areas of banking, including performance assessment, credit evaluation, risk management, customer retention, and fraud detection. However, when we carefully review the above work, it is easy to see that the machine learning algorithms used in the above work are mostly explanatory models, which are used to verify the causal relationships between observable variables in the theory. Unlike the previous work mentioned above, our work in this paper is based on predictive analysis, which has appeared less frequently in empirical studies of finance and banking. The method proposed in this paper does not assume a causal relationship between variables, and most of the models that fit well do not assume a specific functional form between variables (e.g., linear relationship, U-shaped relationships, and exponential relationships), and thus predictive models are able to uncover more complex patterns in the datasets.

3. Methods

3.1. Variable Selection. This paper studies the performance of China's provincial rural banks, that is, provincial rural banks represent the regional heterogeneity of rural banks. Fukuyama and Weber [5] used a two-stage network model including good and bad output to evaluate the performance of Japanese banks. They use labor, physical capital, and financial equity capital to produce loans and securities investments and use deposits as intermediate output.

In order to evaluate the performance of rural banks in different provinces, this paper selects 30 provincial rural banks across the country except Tibet as the research object and uses 4 years of data to evaluate the productivity growth and decomposition efficiency indicators of provincial rural banks in China. According to the concept in the literature [6–8], the input variables are capital and employees based on cost and the ideal output variable is profit based on revenue. In addition, this paper studies the dynamic development and risk control of rural banks in China and incorporates carry-over activities and negative externalities into the study, as shown in Table 1.

Banks use capital and human resources to make profits. Bank deposits are considered as a special resource because banks strive to attract deposits and use them as a positive indicator of performance evaluation. At the same time, they use these deposits to earn future profits. In DEA banking literature, deposit is a controversial topic. Compared with other input-output variables, deposits have the characteristics of dynamic variables. Therefore, rural bank deposits are defined as a carry-over variable. From a more comprehensive analysis, nonperforming loans (NPLs) represent bad debt risks, and there is an inevitable symbiotic relationship between bad debt risks and profits. Therefore, the nonperforming loans of rural banks are defined as the nonperforming output of rural banks.

3.2. Model Construction and Algorithm

3.2.1. Malmqusit Index Calculation. The provincial rural bank is defined as the decision-making unit of the performance evaluation of the rural bank, and it is the research object of the performance evaluation of the rural bank. In period t, provincial rural banks (DMUs) use input X and carry-over activity z to produce ideal output Y_d and bad output Y_u. Carry forward activity connection time periods $t-1$, t, and $t+1$. The variables of input, output, and carry-over activities have regional heterogeneity.

In the traditional dynamic DEA model, (X_t, Y_t) and (X_{t+1}, Y_{t+1}) are separately dealt with for obtaining catch-up effect and frontier-shift effect. However, Tone and Tsutsui (2010) introduced the carry-over into the dynamic model, called dynamic SBM (DSBM). This paper basically follows the dynamic SBM thinking. To estimate the frontier functions, upon which we compute the nonoriented measures of the efficiency, we deal with n DMUs ($j = 1, \ldots, n$) over period t ($t = 1, \ldots, T$). Using period t as a benchmark, DMUs produce s outputs ($i = 1, \ldots, s$) using m inputs ($i = 1, \ldots, m$). Moreover, we define r links ($i = 1, \ldots, r$) as carry-over activities between two consecutive periods. Then, we can obtain the pure technical efficiency for DUM j in period t as follows:

$$\rho_v^* = \min \frac{1 - (1/(m+r))\left(\sum_{i=1}^{m}\left(S_{it}^-/x_{iot}\right) + \sum_{i=1}^{m}\left(S_{ilt-1}^-/z_{iot-1}\right)\right)}{1 + (1/(s+r))\left(\sum_{i=1}^{s}\left(S_{it}^+/y_{iot}\right) + \sum_{i=1}^{m}\left(S_{ilt}^+/z_{iot}\right)\right)}, \tag{1}$$

where x_{ijt} and y_{ijt} are the inputs and outputs of DMU_j at period t, respectively, and we define z_{ijt} as links. S_{it}^-, S_{it}^+, S_{ilt-1}^-, and S_{ilt}^+ are slack variables denoting, respectively, input excess, output shortfall, link excess, and link shortfall.

Solving the above program for each DMU, we can obtain ρ_c^*, which means the variable returns to scale case. For the constant returns to scale case ρ_v^*, we only need delete the restriction $\sum_{j=1}^{n}\lambda_j = 1$ in the above model. Then, we can decompose the technical efficiency (TE) into scale efficiency (SE) and pure technical efficiency (PuTE) as

$$\begin{aligned} \overline{\mathrm{TE}} &= \rho_c^*, \\ \overline{\mathrm{PuTE}} &= \rho_v^*, \\ \overline{\mathrm{SE}} &= \rho_c^*/\rho_v^*. \end{aligned} \tag{2}$$

Finally, using the above formula, we can decompose the sources of catching-up effect as

$$\mathrm{CU} = \mathrm{TEC} = \mathrm{PUTC}^*\mathrm{SEC}. \tag{3}$$

According to the above, we can decompose the sources of frontier-shift effect as

Table 1: Data description.

Index	Capital stock	Staffs	Deposit	Profit	NPLR
Mean	71.25	21976	1210.84	1449.96	12.43
SD	61.71	15232	1163.65	1361.5	18.3
Min	0.58	2087	28.26	33.01	−29
Max	234.15	60896	5940.71	6957.89	104.29

Note. NPLR: nonperforming loan ratio.

$$\mathrm{FS} = \mathrm{DPC}^{*}\mathrm{TPC}. \tag{4}$$

In conclusion, we decomposed the dynamic Malmqusit model as

$$\overline{M}(x, y, z) = \overline{\mathrm{TEC}} \cdot \mathrm{DTPC} = \overline{\mathrm{SEC}} \cdot \overline{\mathrm{PuTC}} \cdot \mathrm{DTPC}. \tag{5}$$

In the clustering part, we use hierarchical clustering, gradient boosting regression tree algorithm, and other related algorithms to further cluster the above index results. The hierarchical clustering uses the BIRCH algorithm. This algorithm is mainly used when the amount of data is large and the data type is numerical. We use the adaptively reduced step size gradient boosting regression tree algorithm proposed in this paper to optimize, so as to make the clustering effect better.

3.2.2. Adaptively Reduced Step Size Gradient Boosting Regression Tree. The gradient boosting regression tree algorithm is widely used in clustering research in the financial field. The existing gradient boosting regression tree method has certain shortcomings. Firstly, the existing methods rely too much on data quality, which makes us often unable to achieve the desired prediction accuracy in actual modeling. Secondly, the existing methods require careful adjustment of parameters, and the training time may be relatively long. Finally, the improvement effect of existing methods is relatively limited.

Next, we will introduce the adaptively reduced step size gradient boosting regression tree algorithm. In the gradient boosting regression tree algorithm, the reduction step size is fixed, and it is determined as a parameter when starting to train the model. We now analyze the loss function of the model. Let $H_j(x)$ be the integrated learner of the first j residual trees, let $h_{j+1}(x)$ be the $j+1$ weak learner, and the learning step is λ. The probability of each training sample being selected as a random subsample is $1/n$, so the loss function can be defined as

$$L\left(y, H_j(x) + \lambda h_{j+1}(x)\right) = \sum_{i=1}^{n} \frac{1}{n}\left(y_i - \left(H_j(x_i) + \lambda h_{j+1}(x_i)\right)\right)^2. \tag{6}$$

Given $H_j(x)$ and $h_{j+1}(x)$, in order to find the corresponding reduction step λ when the loss is the smallest, let the loss function take the derivative of λ and make the derivative equal to 0, we can get

$$\frac{\partial L}{\partial \lambda} = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i\left(H_j(x_i) + \lambda h_{j+1}(x_i)\right)\right)h_{j+1}(x_i) = 0. \tag{7}$$

Then, we have

$$\lambda = \frac{\sum_{i=1}^{n}\left(y_i - H_j(x_i)\right)h_{j+1}(x_i)}{\sum_{i=1}^{n} h_{j+1}^2(x_i)}. \tag{8}$$

Therefore, the reduction step size can be automatically updated with the current learning result to adapt to the minimization of the function.

Then, we can write the improved gradient boosting regression tree Algorithm 1 steps as follows.

4. Results

4.1. Experimental Methods and Processes. The experimental data in this paper are the four-year data of 30 provincial rural banks except Tibet, including deposits, capital stock, employees, profits, and nonperforming loan rates. The five efficiency indexes decomposed by the Malmquist index method are SuEC, PuTC, SEC, DPC, and TPC. Taking Yunnan Province as an example, these five indicators are shown in Table 2:

The experimental process of this paper is shown in Figure 1.

The classification part is to divide the rural banks in 30 provinces into several groups, so that the above groups can be divided into different performance categories based on the characteristics of the efficiency of rural banks.

In clustering, we use the BIRCH algorithm and the algorithm proposed in this paper, respectively. Use the original classification results of 30 provinces as a reference to check the accuracy of clustering by these two algorithms.

4.2. Clustering of Rural Bank Performance

4.2.1. BIRCH Clustering. As shown in Figure 2, when using the BIRCH algorithm to classify existing data, we get a total of six groups of results. Since the cluster feature tree has a limit on the number of cluster features of each node, the clustering result may be different from the real category distribution. In addition, the algorithm has a poor clustering effect on high-dimensional feature data.

4.2.2. Gradient Boosting Regression Tree Clustering. As shown in Figure 3, when we use the gradient boosting regression tree algorithm, we get seven groups of provincial banks. The accuracy of the algorithm is higher, the generalization ability is stronger, and the classification result is basically consistent with the original reference.

4.2.3. Performance Type. According to the cluster analysis result and the character of decomposed efficiency in Chinese rural banks, we merge special groups for analysis, such as Group 4 and Group 6 as TPEI (traditional pure economic improved type) and Group 2, Group 5, and Group 7 as SuECI (sustainable efficiency change improved type). This grouping sounds more realistic and good to empirical analysis, so we distinguish Chinese rural banks into four type of performance as shown in Table 3.

Input:
Training samples $T = \{(x_i, y_i) = (x_{i1}, ..., x_{ip}, y_i) | i = 1, ..., n\}$
Residual tree training times is M, random sampling rate is rate, complexity parameter is cp.
Training steps:
Initialize training samples $T_1 = T$, *where* $y_j = (y_1, y_2, ..., y_n)$, *reduce step size* $\lambda_1 = 0.01$,
FOR $j = 1, 2, \ldots, M$
(1) *From* T_j *without replacement, repeat the subsample with a random ratio of rate as the training sample of the current regression tree.*
(2) *Based on the complexity parameter cp, train the* $j-$th *residual tree model* $h_j(x)$ *on the current training sample.*
(3) Update reduction step $\lambda_j = \sum_{i=1}^{n} (y_i - H_{j-1}(x_i))/h_j(x_i)$.
(4) *Give the predicted value* $\widehat{y}^j = (\widehat{y}_1^j, \widehat{y}_2^j, \ldots, \widehat{y}_n^j)$ *of the training sample* T_j *on* $h_j(x)$.
(5) *Update the output variable value* $y^{j+1} = y^j - \lambda_j \widehat{y}^j$ *on the training sample* T_j.
END FOR
Output: improved gradient boosting regression tree model $H(x) = h_M(x) + \sum_{j=1}^{M-1} \lambda_j h_j(x)$.

Algorithm 1: Gradient boosting regression tree with an adaptively reduced step size.

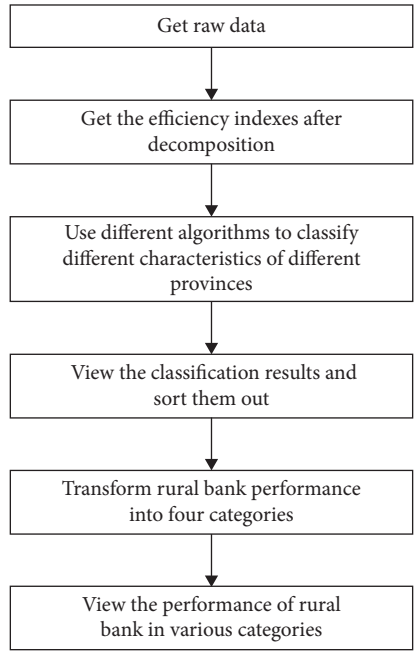


Figure 1: The experimental process outline.

Types (I) and (II) rural banks perform lower than type (III). While from the sustainable development viewpoint, types (I) and (II) belong to potential banks and type (III) exists implicit crisis. We refer to type (III) as cash cows in Boston matrix. Rural banks of type (IV) are diverse. However, the unified advantage of sustainable efficiency makes them stand out as part of a sustainable development strategy. Hereafter, we analyze the four types.

4.2.4. Rural Bank Performance in Inland Areas. Most DPCL banks are located in inland areas in China. The performances of rural banks seriously lag behind other three type banks. The main characteristic is that DPC is the only bottleneck that constrains their performance. From purely a profit viewpoint, PuTC is on the efficient frontier and TPC improves productivity growth. This suggests allocation of inputs and desirable outputs are effective and their quality growths are positive. However, the undesirable outputs and links are ineffective. That is to say, these banks aim at pursuing short-term profit and ignore long-term sustainable profit.

As shown in Figure 4, the Gansu rural bank keeps pure profit indexes effective. Meanwhile, its highly sustainable efficiency changes keep its performance rank the top seven in China. This suggests that, at the primary period of sustainable development, incorporating sustainable method into performance estimation makes greater progress. Above all, though DPCL performance lags behind others, it has the only bottleneck of carry-over activity (deposit). This presents challenges as well as opportunities.

4.2.5. Rural Bank Performance in Coastal Areas. The SuTECL banks are located in the coastal panhandle of the east area, which includes seven provinces. Besides the coastal

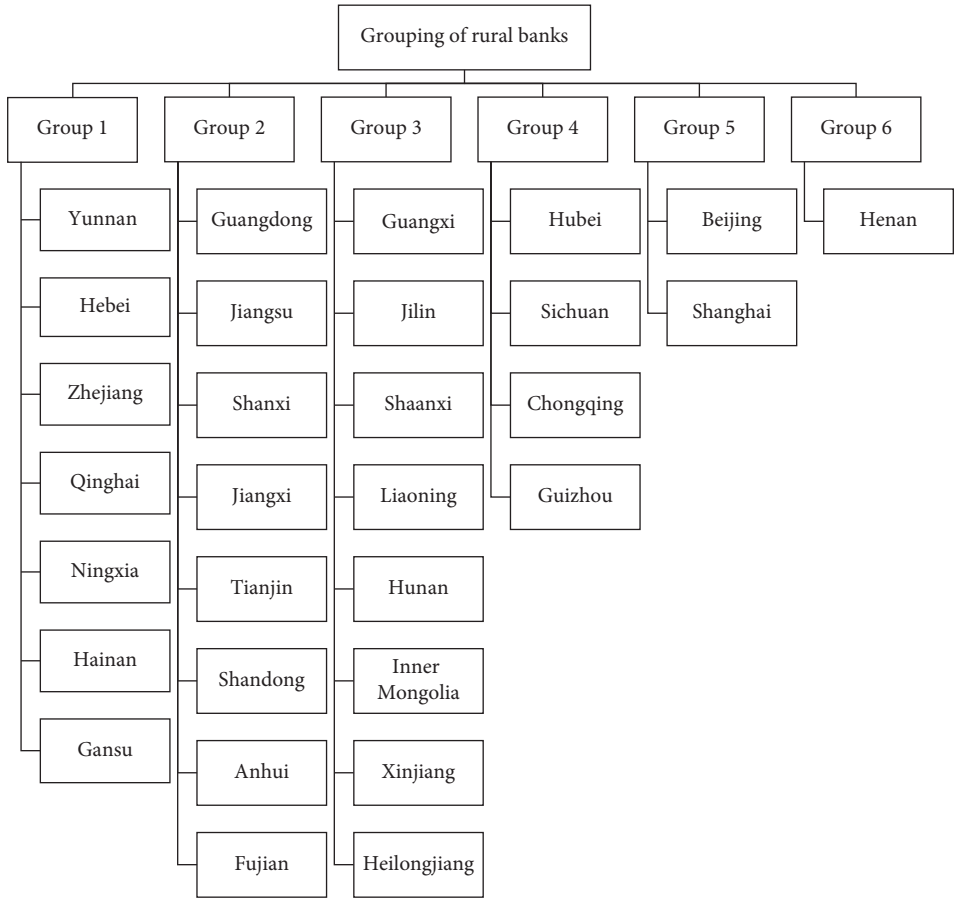


Figure 2: The clustering result of Birch algorithm.

panhandle of the east area, Shanxi rural bank also belongs to the SuTECL. The performances of these rural banks are in the bottom half in Chinese rural banks. The main characteristics are that TPC is the only benefit, and lower PuTC and medium below SuEC in SuTECL type banks limit the productivity growth of these banks. This suggests the local developed economy drives the improvement of performance. However, allocation of inputs and desirable outputs loses the customary advantage in the eastern area. That is to say, this is a big challenge because these banks ignore the basic control of factor efficiency.

As shown in Figure 5, Anhui rural bank is the only one whose DPC is effective. Although it is ineffective from a pure profit perspective, this bank focuses on a sustainable development strategy. So, the control of link and undesirable output improves its performance and puts it in the top two of this type. The Anhui rural bank is lower than that of Fujian. However, with the viewpoint of the sustainable efficiency change of Anhui rural bank, its performance will exceed Fujian's in the near future.

4.2.6. Rural Bank Performance in Central Areas. The banks of TPEI are located in the panhandle of northern and central regions of China as *T* sharp, including five provinces in northern China. Besides that, Hubei and Hunan rural banks also belong to TPEI as shown in Figure 6. The performances of these rural banks are higher and show smooth fluctuation. The main characteristics are that DPC is lower, and the performances of PuTC and TPC improve together. This suggests it has advantages from a purely technical viewpoint. The performances of these banks are in the leading position among Chinese rural banks. However, it is a big challenge to this type since lower DPCs in these banks mean ignorance of the deposit scroll effect in the long term. It will be difficult for this type of bank to keep its predominance if it continues to pursue short-term profit. This will also have a series of drawbacks.

As shown in Figure 6, the number of rural banks in TPEI is one-third of 30 banks in China. So for Chinese rural banks, it is still a long way to control carry-over activity (deposit) and undesirable output (NPLR) and the situation is severe. Rural banks in Xinjiang, Liaoning and Jilin are effective from a sustainable development strategy viewpoint. The effective situation means these banks have already focused on developing a sustainable dynamic strategy, especially deposit and NPLR control. These type banks are referred to as cash cows in Boston matrix. So, using the profit advantage, if it gradually transfers the focus into sustainable development strategy, it will be in the leading position of China.

4.2.7. Rural Bank Performance in Municipality Areas. The SuECI banks have the advantage of being new, and these type banks include Henan bank and the three municipality

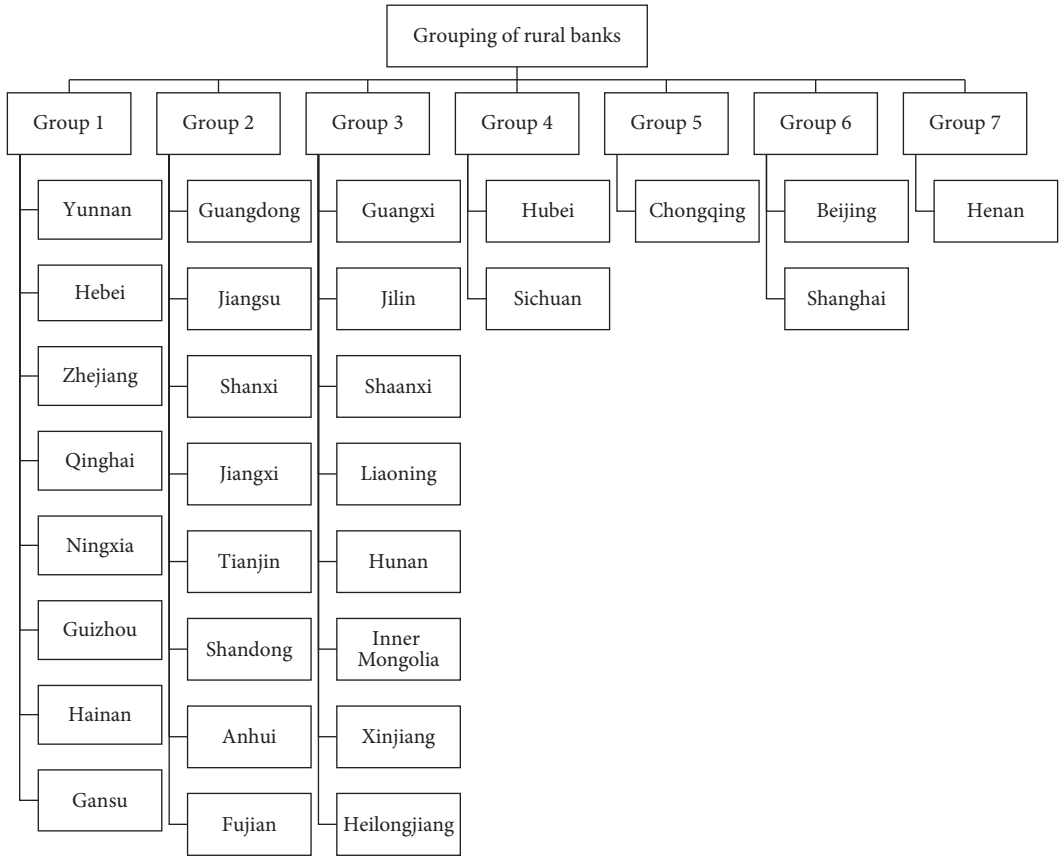


Figure 3: Gradient lift regression tree clustering.

Table 2: Five indicators of Yunnan province.

Inland	DSTFP	SuEC	PuTC	SEC	DPC	TPC
Yunnan	0.99	1.00	1	1	0.89	1.12

Table 3: Four type of performance of Chinese rural banks.

No.	Abbreviation	Norm
(I)	DPCL	Dynamic progress limited type
(II)	SuTECL	Sustainable technical efficiency changes limited type
(III)	TPEI	Traditional pure economic improved type
(IV)	SuECI	Sustainable efficiency changes improved type

banks in Chongqing, Beijing, and Shanghai. The performances of these rural banks differ significantly. The main characteristics are that SuEC and TPC are higher. The characteristics mean that the performances of SuECI banks have benefited from local economic advantages and sustainable development strategy. This is an opportunity for great performance improvement because of the sustainable advantage.

As shown in Figure 7, Chongqing is the youngest municipality in China. The sustainable dynamic performance (DSTFP) is the lowest among Chinese rural banks. The main reason is its low profit efficiency. The scroll of deposit and the control on NPLR have advantage in rural banks of China. This presents challenges as well as opportunities. The Beijing and Shanghai rural banks are in developed areas in China. Strong local economies there improve the performance of rural bank. However, the control of deposit is a drawback, especially in Beijing. The drawback means sustainable development requires a qualitative leap after quantity accumulates. Otherwise, it is difficult to continue performance improvements. The performance in Henan rural bank is the best from the viewpoints of both sustainability and allocation. This proves that Henan rural bank seizes the opportunity even if it does not have a strong economic backdrop. That is to say, at the primary period of sustainable development, incorporating sustainable methods into performance management can improve the productivity growth greatly.

4.3. Contrastive Analysis of Rural Bank Performance. After analysing the four types of rural banks in my country, the results of the model before and after using machine learning technology are compared. This can more clearly show our contribution to empirical analysis.

Based on the above model, we compared the total factor productivity of China's rural banking industry. On the whole, the use of machine learning technology has a more obvious positive effect on bank performance evaluation, especially for high-efficiency banks. It refers to provinces that are purely economically efficient, ignores sustainable

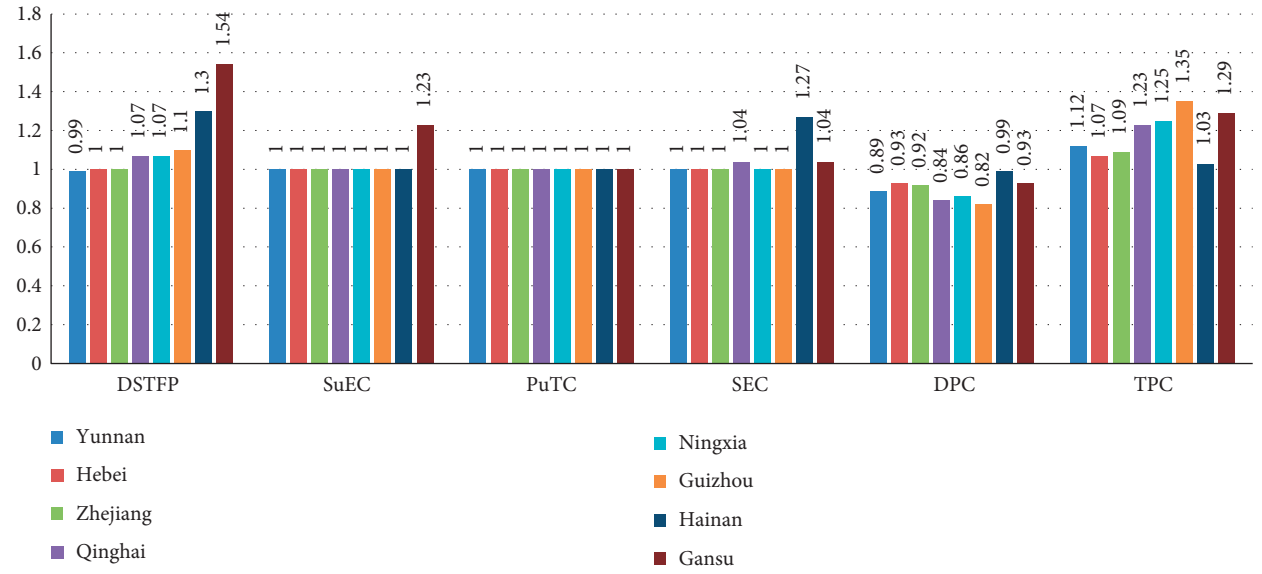


Figure 4: Decomposed efficiency indexes of DPCL.

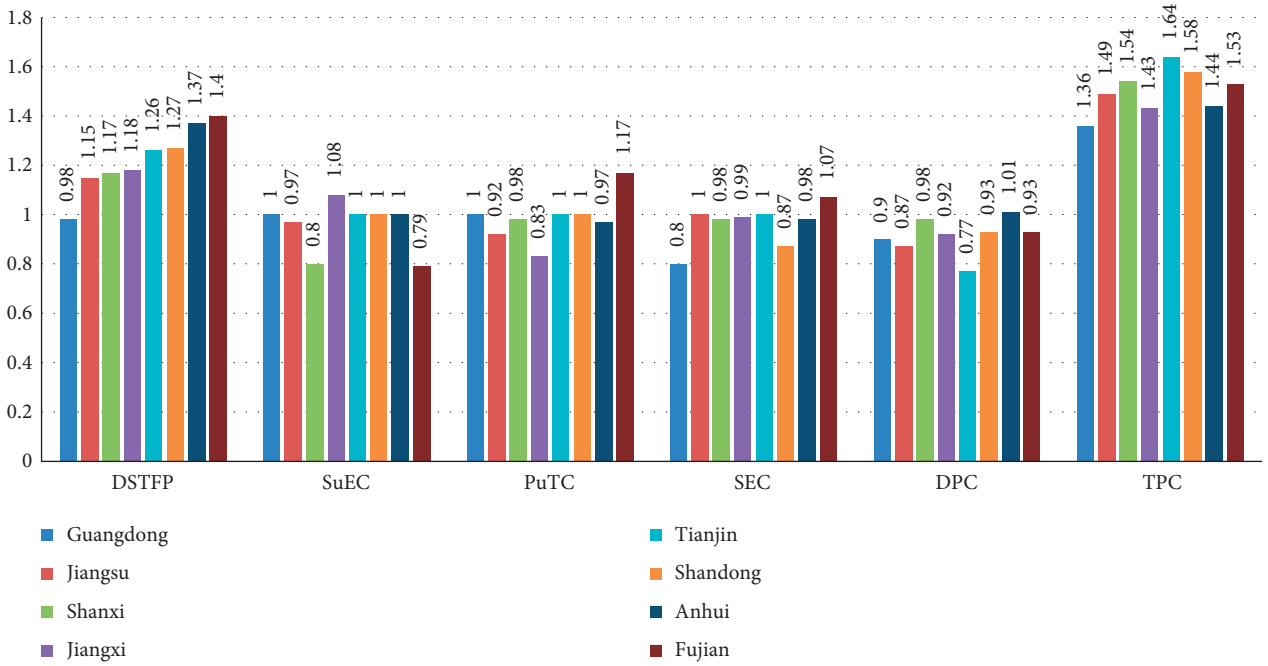


Figure 5: Decomposed efficiency indexes of SuTECL.

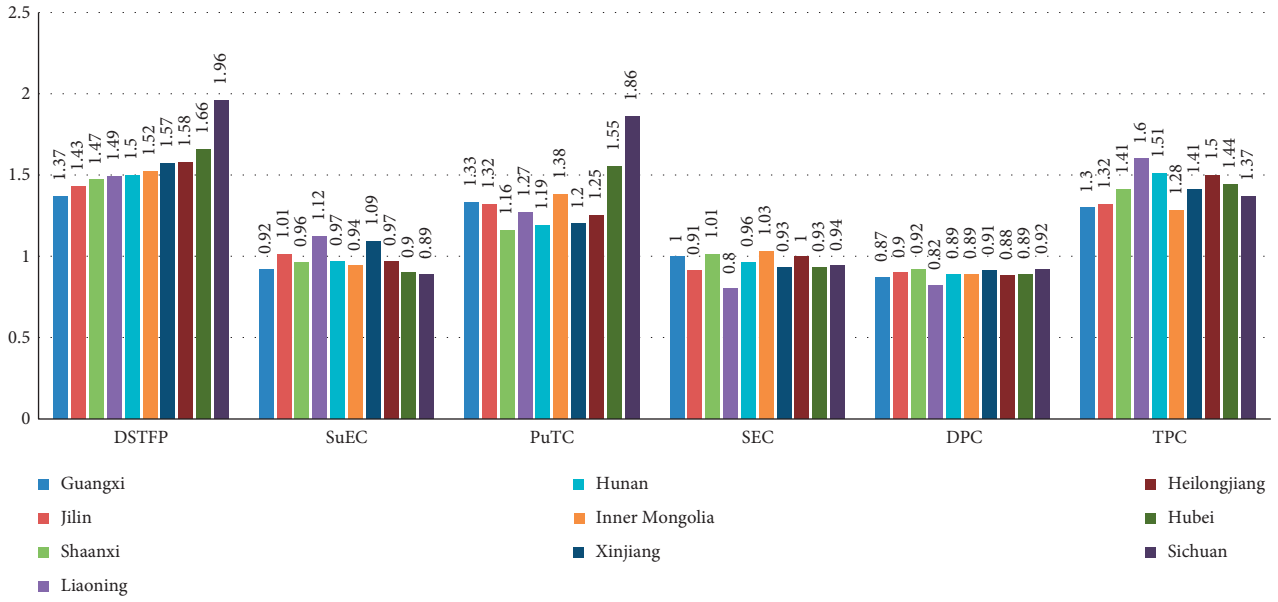


Figure 6: Decomposed efficiency indexes of TPEI.

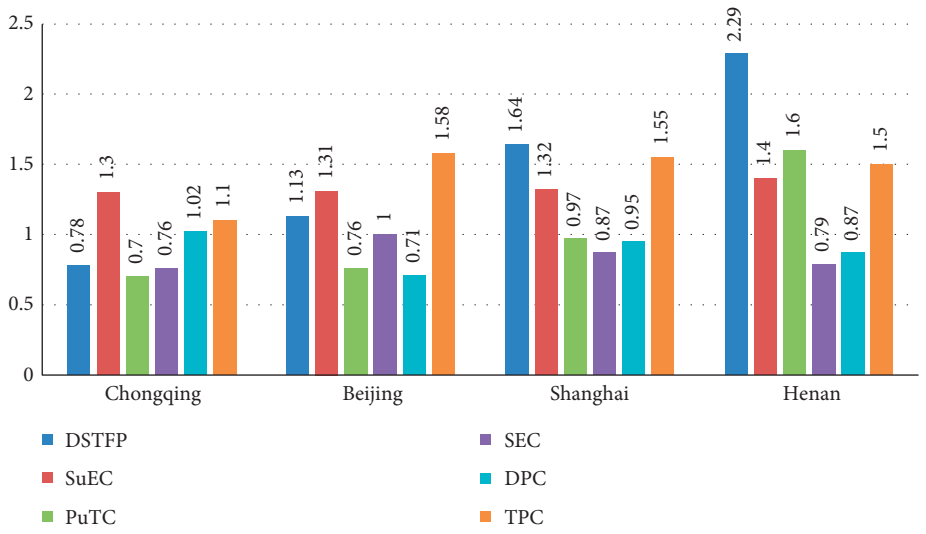


Figure 7: Decomposed efficiency indexes of SuECI.

development, and emphasizes short-term development. Among these banks, the rolling effect of efficiency and loan interest rates restricts the sustainable development of rural banks. As an inefficient bank in a purely economic sense, sustainable dynamic efficiency has a positive impact on its performance. For example, Xinjiang Rural Bank has a good performance and sustainable dynamic efficiency has played a positive role. The development model of the region is in good condition and needs attention.

In summary, it is still a process to incorporate sustainable development strategies into the operation and management of rural banks in my country. It can be seen from the above model that the productivity growth of rural banks is affected by catching up with the effective frontier and shifting from the effective frontier. We have made a comparative analysis of its performance from the perspective of pure economy and sustainable development.

5. Conclusions

In the current performance evaluation works of commercial banks, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics. On the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction. This paper is the first to comprehensively investigate the predictability of multidimensional features on commercial bank performance using boosting regression tree. Aiming at the characteristics of commercial bank data,

this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. Compared to the BIRCH algorithm for classification of existing data, our proposed gradient boosting regression tree algorithm with adaptively reduced step size obtains better classification results. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance.

Based on the hierarchical cluster analysis, the banks in China are divided into four groups: DPCL, SuTECL, TPEI, and SuECI. This paper also summarizes some interesting findings about the productivity growth of various types of rural banks in China, such as SuECI is worthy of attention; TPEI is potentially dangerous. The reason is that although this type of bank has good profit performance, it performs poorly in the evaluation of NPLR.

The follow-up research includes four aspects. First, we will apply external weights to all inputs, links, and outputs [21, 22]. Second, we will incorporate dynamic cost revenue and profit efficiency into our model [23]. Third, we will conduct sensitivity analysis and factor analysis of DSMPI [24]. Fourth, we will apply resampling methods, such as bootstrap techniques, to estimate the performance.

Data Availability

All data used in this study can be made available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] W. Bolt and D. Humphrey, "Bank competition efficiency in Europe: a frontier approach," *Journal of Banking & Finance*, vol. 34, no. 8, pp. 1808–1817, 2010.

[2] J. C. Paradi, S. Rouatt, and H. Zhu, "Two-stage evaluation of bank branch efficiency using data envelopment analysis," *Omega*, vol. 39, no. 1, pp. 99–109, 2011.

[3] D. U. A. Galagedera, I. Roshdi, H. Fukuyama, and J. Zhu, "A new network DEA model for mutual fund performance appraisal: an application to U.S. equity mutual funds," *Omega*, vol. 77, pp. 168–179, 2018.

[4] S. Cheng, R. Lu, and X. Zhang, "What should investors care about? Mutual fund ratings by analysts vs. Machine learning technique," *Machine Learning Technique*, 2020.

[5] H. Fukuyama and W. L. Weber, "A slacks-based inefficiency measure for a two-stage system with bad outputs," *Omega*, vol. 38, no. 5, pp. 398–409, 2010.

[6] T. Kaoru, M. Tsutsui, and D. E. A. Dynamic, "A slacks-based measure approach," *Omega*, vol. 38, pp. 145–156, 2010.

[7] T. Kaoru and M. Tsutsui, "Network DEA: a slacks-based measure approach," *European Journal of Operational Research*, vol. 197, pp. 243–252, 2009.

[8] T. Kaoru and M. Tsutsui, "Dynamic DEA with network structure: a slacks-based measure approach," *Omega*, vol. 42, no. 1, pp. 124–131, 2014.

[9] V. Ravi, H. Kurniawan, P. N. K. Thai, and P. R. Kumar, "Soft computing system for bank performance prediction," *Applied Soft Computing*, vol. 8, no. 1, pp. 305–315, 2008.

[10] S.-W. Kumar, Y.-R. Shiue, S.-C. Chen, and H.-M. Cheng, "Applying enhanced data mining approaches in predicting bank performance: a case of Taiwanese commercial banks," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11543–11551, 2009.

[11] M. D. Cheng and F. Pasiouras, "Assessing bank efficiency and performance with operational research and artificial intelligence techniques: a survey," *European Journal of Operational Research*, vol. 204, no. 2, pp. 189–198, 2010.

[12] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, "Predicting bank insolvencies using machine learning techniques," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1092–1113, 2020.

[13] H. H. Le and J.-L. Viviani, "Predicting bank failure: an improvement by implementing a machine-learning approach to classical financial ratios," *Research in International Business and Finance*, vol. 44, pp. 16–25, 2018.

[14] T.-H. Chen, "Do you know your customer? Bank risk assessment based on machine learning," *Applied Soft Computing*, vol. 86, p. 105779, 2020.

[15] R. E. Turkson, E. Y. Baagyere, and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," in *Proceedings of the 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, pp. 1–7, Shenzhen, China, September 2016.

[16] P. Appiahene, Y. M. Missah, and U. Najim, "Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks," *Advances in Fuzzy Systems*, vol. 2020, Article ID 8581202, 12 pages, 2020.

[17] H. Erdal and İ. Karahanoğlu, "Bagging ensemble models for bank profitability: an emprical research on Turkish development and investment banks," *Applied Soft Computing*, vol. 49, pp. 861–867, 2016.

[18] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: a literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[19] P. Hájek, "Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns," *Neural Computing and Applications*, vol. 29, no. 7, pp. 343–358, 2018.

[20] P. S. Patil and N. V. Dharwadkar, "Analysis of banking data using machine learning," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 876–881, Tirupur, India, February 2017.

[21] A. J. Ikechukwu, "Assessment of organizational performance of private manufacturing companies: the impact of supply chain management responsiveness," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 26–44, 2019.

[22] W. Ghodbane, "Corporate social responsibility and performance outcomes of high technology firms: impacts on open innovation," *Journal of System and Management Sciences*, vol. 9, no. 4, pp. 29–38, 2019.

[23] I. Doina, "POPESCU, sebastian-ion CEPTUREANU, Adriana ALEXANDRU, eduard-gabriel CEPTUREANU, relationships between knowledge absorptive capacity, innovation performance and information technology. Case study: the Romanian creative industries SMEs," *Studies in Informatics and Control*, vol. 28, no. 4, pp. 463–476, 2019, ISSN 1220-1766.

[24] B. Lalic, M. Delic, N. Simeunovic, N. Tasic, and S. Cvetkovic, "The impact of quality management purchasing practices on purchasing performance in transitional economies," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 3, pp. 815–822, 2019.

Hindawi
Complexity
Volume 2021, Article ID 6666619, 12 pages
https://doi.org/10.1155/2021/6666619

Research Article

An Improved Integrated Clustering Learning Strategy Based on Three-Stage Affinity Propagation Algorithm with Density Peak Optimization Theory

Limin Wang,[1] Wenjing Sun,[2] Xuming Han,[3] Zhiyuan Hao,[4] Ruihong Zhou,[1] Jinglin Yu,[1] and Milan Parmar[2]

[1]*School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou 510520, China*
[2]*School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, Jilin, China*
[3]*College of Information Science and Technology, Jinan University, Guangzhou 510632, China*
[4]*School of Management, Jilin University, Changchun 130022, Jilin, China*

Correspondence should be addressed to Xuming Han; hanxvming@163.com and Zhiyuan Hao; 15391910163@163.com

Received 9 November 2020; Accepted 22 December 2020; Published 7 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

To better reflect the precise clustering results of the data samples with different shapes and densities for affinity propagation clustering algorithm (AP), an improved integrated clustering learning strategy based on three-stage affinity propagation algorithm with density peak optimization theory (DPKT-AP) was proposed in this paper. DPKT-AP combined the ideology of integrated clustering with the AP algorithm, by introducing the density peak theory and k-means algorithm to carry on the three-stage clustering process. In the first stage, the clustering center point was selected by density peak clustering. Because the clustering center was surrounded by the nearest neighbor point with lower local density and had a relatively large distance from other points with higher density, it could help the k-means algorithm in the second stage avoiding the local optimal situation. In the second stage, the k-means algorithm was used to cluster the data samples to form several relatively small spherical subgroups, and each of subgroups had a local density maximum point, which is called the center point of the subgroup. In the third stage, DPKT-AP used the AP algorithm to merge and cluster the spherical subgroups. Experiments on UCI data sets and synthetic data sets showed that DPKT-AP improved the clustering performance and accuracy for the algorithm.

1. Introduction

Clustering analysis is an important research direction in the field of data mining. Through analyzing the internal structure information and spatial characteristics of massive data samples, the demand information can be obtained. Based on the advantages in data processing, clustering analysis is widely used in various fields of society. For example, clustering analysis can be used to identify viruses from a large number of virus research data; in artificial intelligence, it can be used for face recognition, fingerprint recognition, and other pattern recognition functions; in financial stock market, clustering analysis could be used to predict stock trends and so on. Therefore, how to improve the clustering method for meeting the research needs with massive data, obtain more accurate clustering results, and meet the demand-oriented user groups in the current society have become a hot issue and a key problem which need to be studied urgently by scholars all over the world [1–4].

In 2007, the American researchers put forward a novel clustering learning method named affinity propagation clustering algorithm in *Science*. The algorithm solves the problem of choosing the initial class representative point in the early stage of clustering. At the same time, there is no need to specify the clustering center, which largely avoids the risk that the algorithm will lead to local optimization due to

the improper selection of parameters. However, the original AP algorithm still has some drawbacks as follows: it is unable to accurately deal with high-dimensional data, it needs to manually set the corresponding parameters, for some specific data types, it cannot accurately identify the data internal structure, and so on.

In this paper, in order to improve the clustering accuracy and clustering performance of the AP for the data with different structures and different sizes, in the AP algorithm, it introduced the density peak clustering theory and k-means algorithm and proposed the three-stage affinity propagation clustering algorithm based on combination optimization of density peak theory and k-means algorithm. The processes are listed as follows:

(1) This paper firstly used the density peak clustering algorithm to obtain the local density ρ and δ values and selected the values of $(\rho^*\delta)$ which are ranked in descending order, and the selection quantity of the $(\rho^*\delta)$ value is k.

(2) The k-means algorithm was used to carry on the secondary processing of data, through the DP algorithm determining the k clustering centers, and the data sample was divided into k subgroups

(3) Through AP clustering the subgroups, the new class label of the center point in a subgroup would be assigned to the other element point in the subgroup.

2. Related Works

With the arrival of the era of big data, the AP algorithm has become a very competitive clustering method in the field of data mining, and the applications of the AP algorithm are implemented in many different fields, for example, scholar E. Graham utilized the AP algorithm to propose a novel unsupervised clustering method in the microbial assemblies field [5]. Wang and Cheng introduced the affinity propagation to resolve the data-driven resource management issue for ultradense small cells [6]. Zhou and Xu combined the AP theory to resolve the issues of segmentation stability in the image segmentation field [7]. Aizpurua and Koutstaal utilized the affinity propagation clustering algorithm to research new index of semantic short-term memory and obtained better progress [8]. At the same time, scholar Chen et al. proposed a novel method for stability-based preference selection based on the AP algorithm [9]. Chinese scholars Zhang et al. extended the AP in a principled way to solve the image clustering problem and proposed the unsupervised image clustering method, which obtained the better result [10]. Ding et al. proposed a derived clustering algorithm for mixed-type data employing fuzzy neighborhood [11]. In the biology field, the scholar introduced the AP into the field of neuroscience data mining [12], etc. Also, in the other fields, a substantial number of scholars combined the affinity propagation clustering algorithm theory to handle the complex issues, including the tumor classification problem [13] and urinary-tract symptoms [14]. Because of the advantages of the AP, the application of the AP was accepted by numerous academics, and they introduced the theory of the AP into their research field to improve their original research results.

At the same time, in the original AP algorithm, there is a very important concept which is the similarity. And, it stipulates the Euclidean distance as the similarity calculation method for any two data samples. However, the Euclidean distance indicates the straight-line distance for any two points in a sample space. In view of the drawback of the Euclidean distance, when the AP algorithm analyzed the data set with intricate data framework, it cannot calculate the relevant precise similarity for the data points and finally obtain the inaccurate clustering result [15].

Given the similarity issue of the AP algorithm, many scholars proposed some different improvement algorithms. For example, Wang et al. altered the structure of the original algorithm to propose a novel self-adaptive affinity propagation clustering algorithm based on density peak theory and weighted similarity (DPWSAP). In the improved algorithm, it constructed a density attribution for the AP. Through weighting the density attribution and distance calculation method, the DPWSAP improved the similarity calculation accuracy, and finally, it obtained more accurate clustering results [16]. Wang et al. utilized the structure similarity to alter the original similarity calculation method to propose an adaptive semisupervised affinity propagation clustering algorithm (SAAP-SS). It started from the perspective of semisupervision, through the structure similarity, to handle a nonlinear, low-rank representation problem, then to improve the similarity calculation for data points, and finally to obtain the better clustering performance [17].

As it is known to all, there are two important parameters in the AP algorithm, including the *preference* and damping factor λ, and each parameter plays a momentous role in the clustering process. The *preference* determines the final clustering numbers of the algorithm; when the value is selected higher, the final clustering number will be greater; also, when the value is selected smaller, the final clustering number will be fewer. For the clustering consequence, the suitable value of *preference* is more important. In view of this parameter, scholar Wang et al. proposed a density propagation-based adaptive multidensity clustering algorithm (DPAM), and the algorithm utilized a density propagation to reduce the impact of the parameter value and achieve the optimal clustering results [18]. Also, for the damping factor λ, the parameter can influence the convergence performance of the AP algorithm. In the clustering process, the suitable value of the damping factor can avoid the local optimal circumstance, in view of the different convergence speed of searching the clustering center in different stages; therefore, the value of the dynamic damping factor is very important. Considering the situation, Wang et al. combined the density peak algorithm and cut-off distance theory through these two theories to control the damping factor and improve the convergence performance of the original algorithm [19]. Wang et al. introduced the gravity concept to propose affinity propagation clustering algorithm based on gravity theory (GAP). GAP constructed a novel clustering method under the physical perspective. On the one hand, it improved the accuracy of similarity calculation for data points;

on the other hand, because of the improvement of algorithm structure, the GAP can control the convergence process of the algorithm well, and it reduced the impact of damping factor and improved the final clustering results [20].

From the above AP application and improved algorithms, we can learn that though the AP algorithm possesses the better application prospect, it owns some defects. The scholars introduced other research theories to improve the AP. However, these improvements have not changed the recognition performance on the data with different structures. They just made the AP obtain the relevant accurate clustering numbers. For the data samples with different shapes and densities, they could not obtain the better clustering result yet. Consequently, in order to improve the clustering accuracy and clustering performance of the algorithm for the data with different structures and different sizes, this paper introduced the DP algorithm and k-means algorithm and used three stages to cluster the data sample, and it improved the accuracy and efficiency compared with the original AP [21–24].

3. Theoretical Basis

3.1. Density Peak Clustering Algorithm. In 2014, the density peak clustering algorithm (DP) was proposed in *Science* , and compared with the early diversity clustering algorithms, the DP arose with a considerable breakthrough, and it greatly improved the performance of clustering algorithm [25]. In the literature [25], there are some merits in determining the final clustering centers. Through introducing the concept of cut-off distance and local density, the DP could apply to analyze the data samples with different types relatively better, including different densities and different shapes. And, the two parameters play the more important role in the clustering process. There are two assumptions making the DP effective [25]:

(1) The cluster center points are embraced by adjacent points with low local density

(2) For the data points with a larger local density, there is a relatively long distance between any two points

In the DP algorithm, a scientific cut-off distance d_c is adopted to calculate their local density ρ for sorting these density values in the descending order as follows [25]:

$$\begin{aligned} d_{ij} &= \operatorname{dist}\left(x_i, x_j\right), \\ \rho_i &= \sum_j \chi\left(d_{ij} - d_c\right), \\ \rho_i &= \sum_{j \neq i} \exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right). \end{aligned} \tag{1}$$

When the algorithm creates the decision graph, there is a formula as follows:

$$\delta_i = \min_{j:\rho_j > \rho_i}\left(d_{ij}\right). \tag{2}$$

Formula (2) represents the minimum distance between the data point i and sample point with higher density. The decision graph is generated according to the ρ value and δ value which are obtained in the definition. As shown in Figure 1 [25], this is the distribution of data point with density size. And, in Figure 2 [25], the data point 10 and data point 1 have relatively high distance and local density at the same time, so they are clustering centers. However, the data points 26, 27, and 28 have relatively high distance, but the local density is smaller, so they are called outliers. For regular data points, the DP algorithm categorizes them into the category of the closest class center, that is, denser than theirs.

3.2. K-Means Algorithm. The k-means algorithm takes k as the parameter and divides n data objects into k classes. The data objects in each class have high similarity, but the similarity between different classes is relatively low. Similarity is calculated by calculating the average value of a data object in a cluster, and the definition of similarity is the key to division. The basic idea of the k-means algorithm is to randomly select k objects as the initial clustering center among n data objects; then, according to the principle of minimum distance, the distance from each data object to the clustering center is calculated and assigned to the nearest cluster. Then, the average value of each cluster is recalculated, and the convergence function is calculated until the center of each cluster no longer changes, and finally, the algorithm is terminated. Otherwise, the above process is repeated. The process of the k-means algorithm is shown in Table 1.

3.3. Affinity Propagation Clustering Algorithm. The core idea of the AP algorithm is to treat all sample points as potential class representative points and to minimize the decision function through the continuous transmission of two kinds of information: *availability* and *responsibility* so that the sample similarity within the cluster is the largest, and the sample similarity between different clusters is the smallest. Assume $\{x_1, x_2, \ldots, x_n\}$ to be a finite data set of the pattern space R_n, where x_i (i could have values of 1, 2, ... , ...) is a point composing of n-dimensional attributes, in a vector space. The similarity between any two samples s (i, k) is measured by a negative Euclidean distance [26] and is shown as follows:

$$S(i, k) = -\left\|x_i - x_k\right\|. \tag{3}$$

In the clustering process of the AP algorithm and before the two important information iterations, it needs to determine the value of parameter *preference*, which is s (k, k). This algorithm considers that the larger the value of the s (k, k), the more likely its corresponding point k is selected as the class representative point. In other words, the number of final clustering classes could be affected by the *preference* value. The affinity propagation clustering algorithm initially assumes that all data points could be chosen as potential class representative points with the same possibility, which is setting all s (k, k) to be the same *preference* value. Different *preference* values could result in different clustering results.

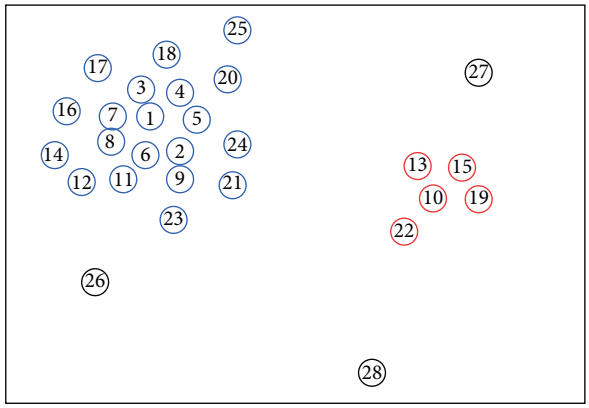


Figure 1: The distribution of the data point with density size.

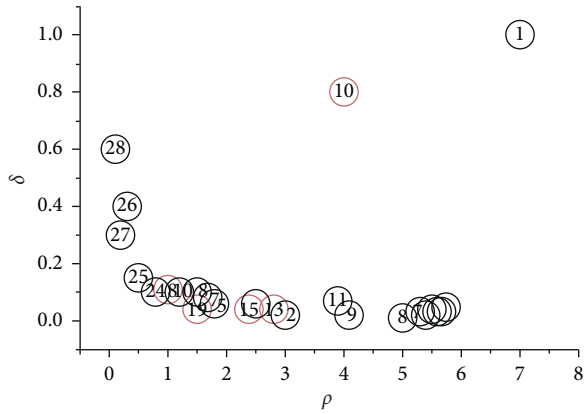


Figure 2: Information transfer between data points.

Generally, the AP algorithm selects the median or minimum value of similarity matrix to be the *preference* value [26].

The AP algorithm has two important information, which are the *responsibility* (r (i, k)) and *availability* (a (i, k)) mentioned above, and each kind of information is a competitive way of different representative points. They propagate continuously between any two data points and finally obtain a more reasonable clustering result. The *responsibility* and *availability* are constantly searched for in order to select suitable class representative points. For any sample point, in any iterative update stage, these two kinds of information together determine a certain sample point as a class representative point and which sample points belong to this class representative point. The iterative process of AP algorithm is actually the process of responsibility and availability alternatively updating the information. *Responsibility* indicates that the data point i sends the information to candidate class representative points k, reflecting the accumulated evidence of point k as cluster center of point i. At this time, there are many data samples competing with k point as the class center representative points of data point i. *Responsibility* is the information matrix which is established to select a final potential clustering center. *Availability* indicates that the candidate class representative point k sends the information to data point i, reflecting the accumulated evidence of the possibility for data point i selecting point k as its cluster center. Also, there are other points selecting the candidate class representative point k as their cluster center, and *availability* is also the information matrix which is established for this competitive mechanism [26].

At the beginning, assuming the value of a (i, k) equal to 0, two information updates are as follows:

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} \left\{0, r\left(i', k\right)\right\}\right\} & i \neq k \\ \sum_{i' s.t. i' \neq k} \left\{0, r\left(i', k\right)\right\} & i = k \end{cases},$$

$$r(i,k) = s(i,k) - \max_{k' s.t. k' \neq j} \left\{a\left(i, k'\right) + s\left(i, k'\right)\right\}. \tag{4}$$

Through the updating iteration process of two kinds of message, *responsibility* and *availability*, between data sample points, the decision matrix E determines k as the final class representative point and is as follows:

$$E(k) = \arg\max_{k} \left(a(i,k) + r(i,k)\right). \tag{5}$$

The whole affinity propagation clustering algorithm can use the computer to calculate the two important similarities quickly and then obtain some reasonable numbers of clustering class. The above formulas determine any data sample point i could be the possible class center point in the case of the point i equal to point k. Also, the algorithm will eventually terminate because two kinds of information, *responsibility* and *availability*, are less than a certain threshold, or the local iteration situation does not change.

On the contrary, an important parameter, which is named as damping factor λ, is introduced in the updating of the affinity propagation clustering algorithm to avert numerical oscillation. During iteration, the renovating results of r (i, k) and a (i, k) can be obtained by computing the previous iteration results in each cycle iteration. Damping factor influences the convergence performance of the AP algorithm. When the number of classes generated by the AP algorithm continuously oscillates during iteration and cannot converge, increasing the damping factor can eliminate this oscillation. The range of damping factor values is [0, 1], and the default value is 0.5. The iteration process is as follows:

$$\begin{aligned} r^{(t+1)}(i,k) &\leftarrow (1-\lambda) r^{(t+1)}(i,k) + \lambda r^{(t)}(i,k), \\ a^{(t+1)}(i,k) &\leftarrow (1-\lambda) a^{(t+1)}(i,k) + \lambda a^{(t)}(i,k). \end{aligned} \tag{6}$$

4. Research Method

In order to improve the clustering accuracy and clustering performance of the algorithm for the data with different types and different sizes, this paper introduced the DP algorithm and K-means algorithm into the AP algorithm and

Table 1: The process of the k-means algorithm.

Step 1: randomly select k objects as the initial clustering center among n data objects
Step 2: according to the principle of minimum distance, the distance from each data object to the clustering center is calculated and assigned to the nearest cluster
Step 3: the average value of each cluster is recalculated, and the convergence function is calculated until the center of each cluster no longer changes
Step 4: when the cluster center does not change, the algorithm is over; otherwise, it will turn to Step 2

used three stages to cluster the data sample [27–32]. And, the stages are as follows:

(1) In the first stage, the clustering center point was selected by density peak clustering. Because the clustering center is surrounded by the nearest neighbor point with lower local density and has a relatively large distance from other version points with higher density, it could help the k-means algorithm in the second stage avoid the local optimal situation. The process is as follows:

The DP algorithm, firstly, coped with the data sample and obtained the local density ρ and δ values of each point. At the same time, the paper calculated the value of $(\rho^* \delta)$ and ranked the product value in descending order. According to the theory of the DP algorithm, the greater the product value is, the more likely the point is to become a class center. Thus, the paper selected the K-potential class centers by the product value from large to small.

(2) In the second stage, the k-means algorithm was used to cluster the data samples to form several relatively small spherical subgroups. Each subgroup has a local density maximum point, which is called the center point of the subgroup. The process is as follows:

Assume $P = \{p_1, p_2, p_3, \ldots, p_n\}$ is the data point set and $G = \{g_1, g_2, g_3, \ldots, g_k\}$ is the K subgroups which are obtained by the k-means algorithm. The value of K is from the first stage, and the center point of each subgroup is actually the potential clustering center point which is selected in the first stage. $D_K = \{\boldsymbol{D_1}, \boldsymbol{D_2}, \boldsymbol{D_3}, \ldots, D_i, \ldots, \boldsymbol{D_k}\}$ is the distance matrix, which indicates the distance between the elements in the subgroup and the K center points. There are K columns and n_k rows in the D_k. n_k is the number of elements of column j and also is the number of the elements in subgroup g_k. The paper made the distance of any two subgroups as follows:

$$\text{distance}\left(g_i, g_j\right) = \min\left(\min\left(D_i j\right), \min\left(D_j i\right)\right). \tag{7}$$

In formula (7), $D_i j$ is the all value of the column j in distance matrix D_i; $D_j i$ is the all value of the column i in distance matrix D_j, and n_j is the number of elements of column i. This paper used the distance which is defined in formula (7), rather than the distance between the any two center points. The calculation method is to find classes with nonconvex shapes, and formula (7) could provide more information about the compactness of two subgroups.

(3) In the third stage, because the AP algorithm is suitable for dealing with spherical data sets, based on this, the paper used the AP algorithm to merge and cluster the spherical subgroups formed in the second stage and finally realized the clustering analysis process of data samples. Experimental results show that the clustering accuracy of the DPKT-AP algorithm is obviously improved, and the clustering effect is better. The process is as follows.

The AP algorithm used the distance between any two subgroups as the similarity calculation method:

$$S(i, j) = \text{distance}\left(g_i, g_j\right). \tag{8}$$

The process of the DPKT-AP algorithm is in Table 2.

5. The Analysis of Simulation Experiment

To test the feasibility and effectiveness of the DPKT-AP algorithm, this paper compared it with the k-means, AP algorithm, and DP algorithm in three UCI data sets and two synthetic data sets listed in Table 3.

For proving the clustering accuracy of the developed DPKT-AP algorithm, this paper selected the three different algorithms which are the k-means, DP algorithm, and AP algorithm to compare with the DPKT-AP algorithm. According to the different densities and the different characteristics of the data sets to verify the clustering accuracy for the improved algorithm, we could use the clustering result to reflect the advantage of the DPKT-AP algorithm. The simulation experiment of the k-means, DP, original AP, and DPKT-AP algorithm was, respectively, tested in 5 different data sets. Comparing the four different clustering results, the following figures are clustering results. The paper could obviously obtain that through the three-stage clustering, and the DPKT-AP algorithm can obtain more accurate clustering numbers.

The subgroup center point of five different data sets is shown Figure 3, and as shown from Figures 4–8, the proposed DPKT-AP algorithm and the DP can aggregate clusters with varying structures and varying densities. The k-means and original AP algorithms cannot obtain the accurate clustering results. Flame and Aggregation belong to different structure data sets; Jain, D1, and D2 belong to different density data sets. For Flame and Aggregation data sets, the DPKT-AP and DP can detect classes of different shapes, and their results are almost the same. The original AP and k-means performed worse on Flame and Aggregation data sets. As for the original AP, no matter how it adjusts its parameters, it cannot find the correct clustering numbers on Aggregation data sets. More importantly, the results obtained by the AP are sensitive to the parameters *Preference* and Damping Factor, and the better results need to be carefully adjusted. For Jain, D1, and D2 data sets, they are made up of clusters of different shapes and densities. The DPKT-AP and DP found the correct clustering numbers on

Table 2: The process of the DPKT-AP algorithm.

Input: similarity matrix S (i, j), cut-off distance d_c value, and initial parameter k
Output: final cluster number, division result $C = \{C_1, \ldots, C_k\}$, and the value of the evaluating indicators
Step 1: select d_c value
Step 2: density peak algorithm is used to calculate the local density ρ value and δ value
Step 3: according to the local density ρ value and δ value, the DP algorithm is used to get the initial clustering center point
Step 4: using the k-means algorithm to iterate the data sample and obtaining the several relatively small spherical subgroups, each subgroup has a local density maximum point, which is called the center point of the subgroup
Step 5: run the AP algorithm to go to the third stage of the clustering process, and use the evaluating indicators to evaluate the effectiveness of the algorithm

Table 3: The different data sets.

Data set	Sample number	Dimension	Class number
D1	**87**	**2**	**3**
D2	**85**	**2**	**4**
Jain	**373**	**2**	**2**
Flame	**240**	**2**	**2**
Aggregation	**788**	**2**	**7**

three data sets and almost obtained the same results. For D1 data set and D2 data set, the original k-means can get the correct clustering number; the AP could obtain the 3 classes and 5 classes, but they could not obtain the accurate sample data points' allocation. For Jain data set, the k-means algorithm could obtain the 2 classes, but the AP obtained the result with 3.

In this paper, in view of improving the clustering accuracy for the AP algorithm, it introduced the DP clustering and k-means algorithm into the original AP algorithm. The DPKT-AP combined the advantages of the DP, which is that the DP algorithm could find the center point quickly, and it has a relative advantage in identifying data with different sizes, densities, and shapes. And, the k-means could analyze the raw data to form spherical subgroups. From the above results, the proposed DPKT-AP algorithm obtains more improvements which are compared with the original AP algorithm. And, these improvements are mainly for the first two stages of the clustering process.

This paper used four different external evaluation methods to analyze the clustering performance of the compared algorithms, including Jaccard coefficient, Rand index, FM index, and $F1$ index. And, there are the following formulas of the four different evaluation methods [16, 20]:

$$M = a + b + c + d = \frac{N(N-1)}{2}. \tag{9}$$

In formula (9), a indicates the amount of data entity pairs which belong to the same class in the clustering results, but belong to different classes in the real structure; b indicates the amount of data entity pairs which belong to the same class in the clustering results and also belong to the same class in the real structure; c indicates the amount of data entity pairs which belong to different classes in the clustering results, but belong to the same class in the real structure; d indicates the amount of data entity pairs which belong to different classes in the clustering results and also belong to different classes in the real structure; N indicates the amount of all data entities [16, 20].

(1) Jaccard coefficient:

$$J = \frac{a}{a + b + c}. \tag{10}$$

(2) Rand index:

$$R = \frac{a + b}{M}. \tag{11}$$

(3) FM index:

$$\mathrm{FM} = \sqrt{\frac{a}{a+b}\,\frac{a}{a+c}}. \tag{12}$$

(4) $F1$ index:

$$\begin{aligned} P &= \mathrm{precision}(i,\ j) = \frac{N_{ij}}{N_i}, \\ R &= \mathrm{recall}(i,\ j) = \frac{N_{ij}}{N_j}. \end{aligned} \tag{13}$$

In formula (13), N_{ij} is the amount of classified i in cluster j; N_j is the amount of cluster j; N_i is the amount of classified i [16, 20]:

$$F1 = \frac{2PR}{P + R}. \tag{14}$$

This paper utilized these evaluation indicator formulas to compare the AP, k-means, DP, and DPKT-AP algorithm. The result showed that the DPKT-AP algorithm is better among the four evaluation indicators. From Tables 4–7, there are evaluation results about the validity of the algorithm. The effectiveness of the algorithm evaluation results are listed in the following tables. From these four evaluation result tables, we can also get that the DPKT-AP algorithm can cluster data more accurately than the k-means algorithm and original AP algorithm, through the combination of the advantages between the k-means algorithm and DP algorithm. When the DPKT-AP processes data of different shapes and densities, it could obtain an apparent improvement of clustering performance, which is compared with the original AP, and it proves the theoretical feasibility for the DPKT-AP.

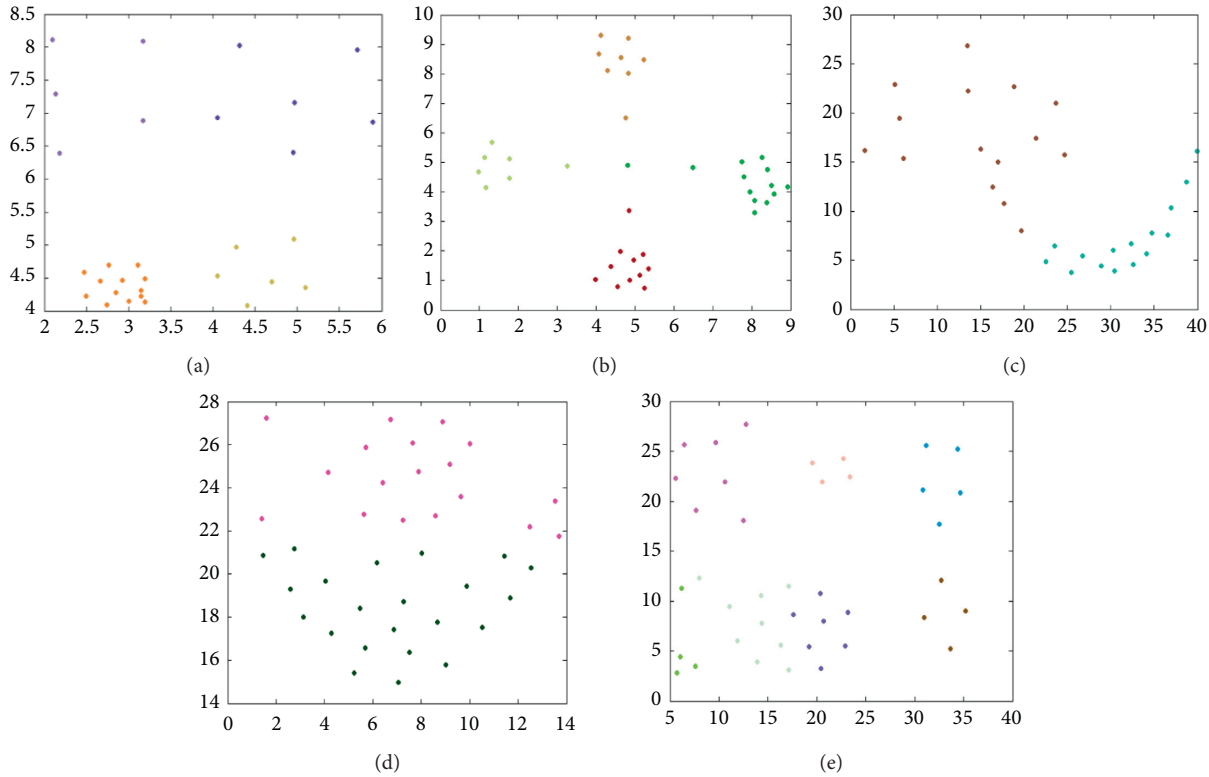


Figure 3: The subgroup center point of five different data sets.

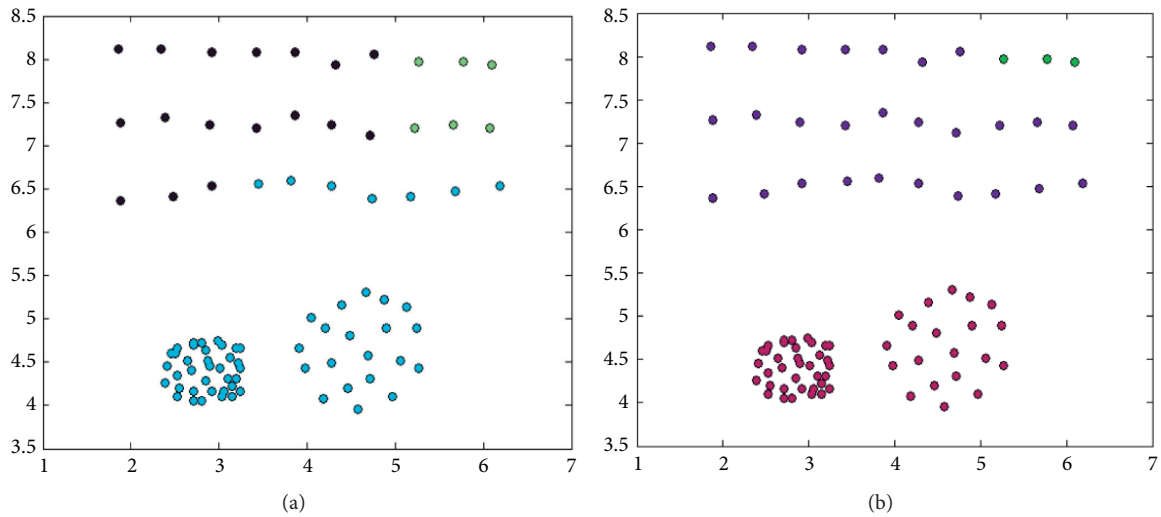


Figure 4: Continued.

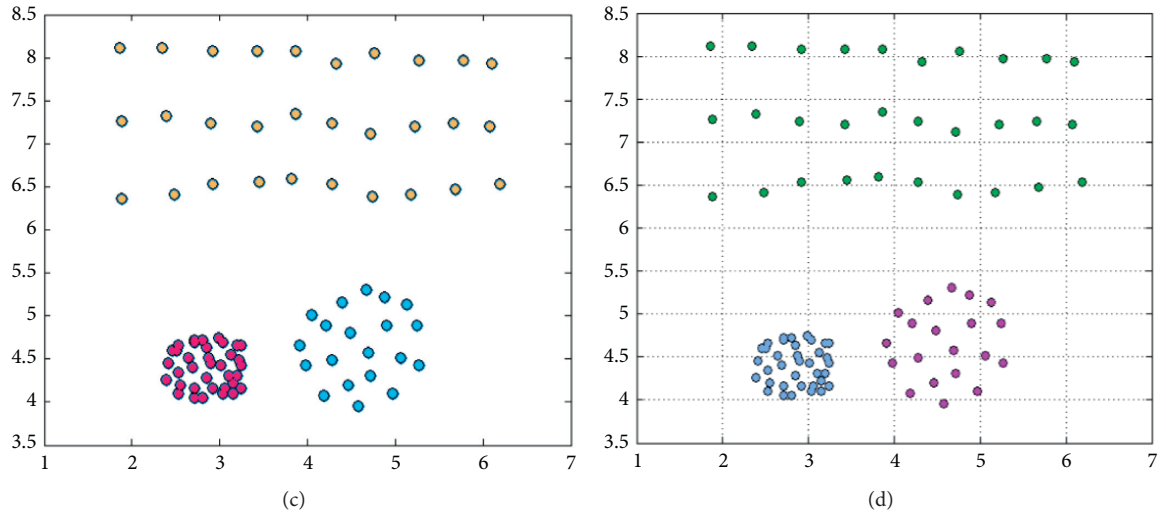


Figure 4: The clustering results of D1 data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

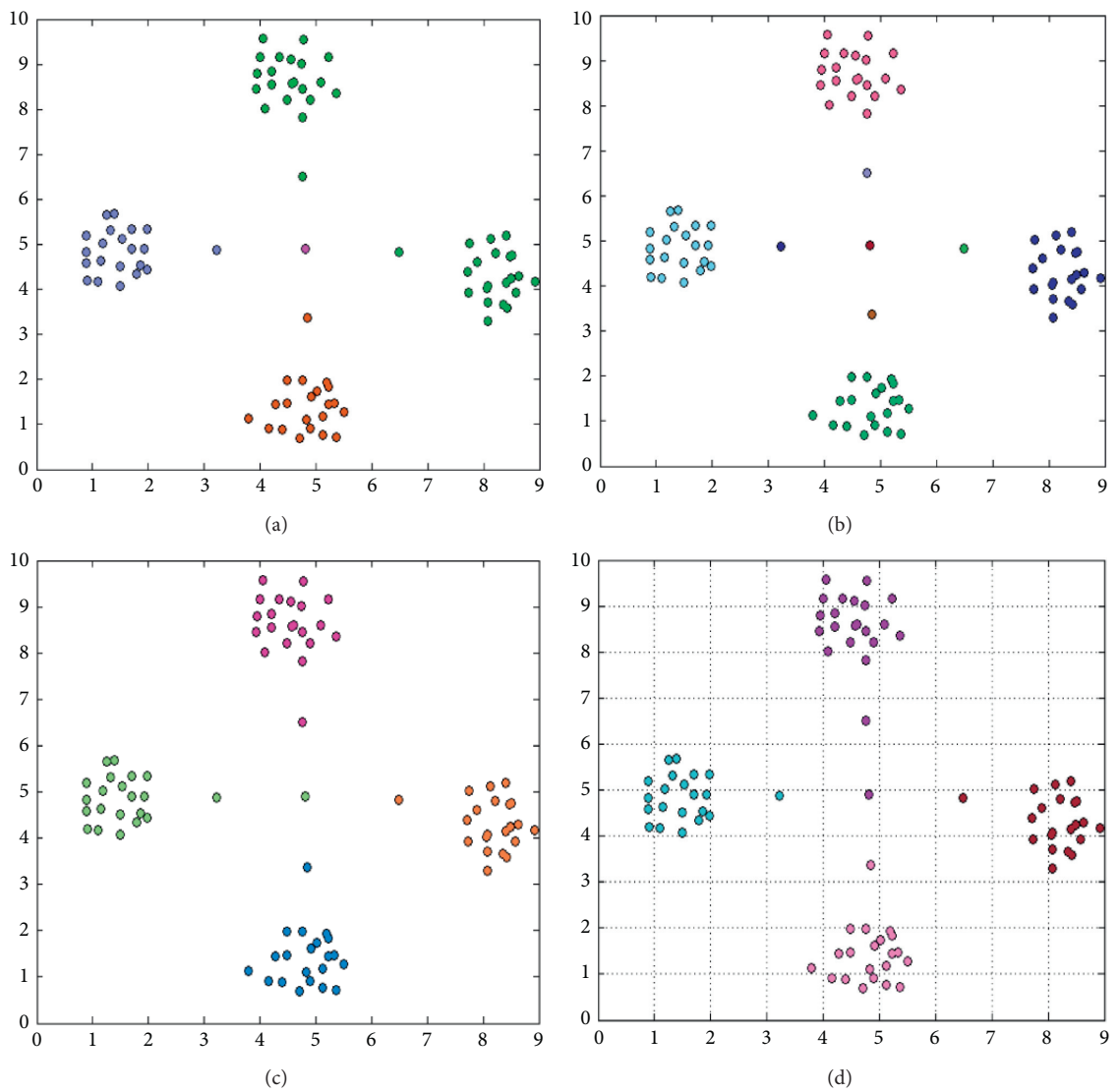


Figure 5: The clustering results of D2 data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

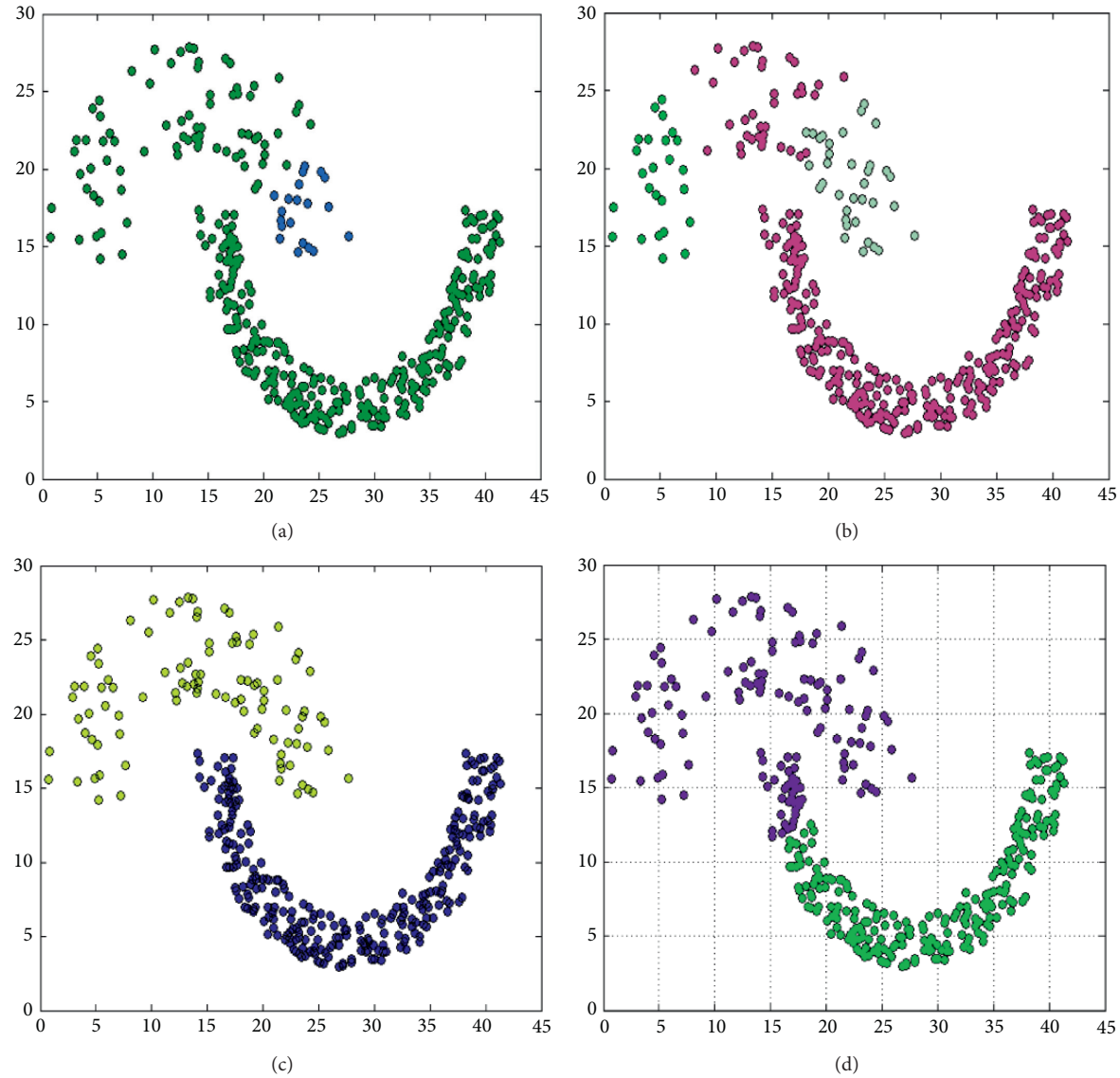


Figure 6: The clustering results of Jain data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

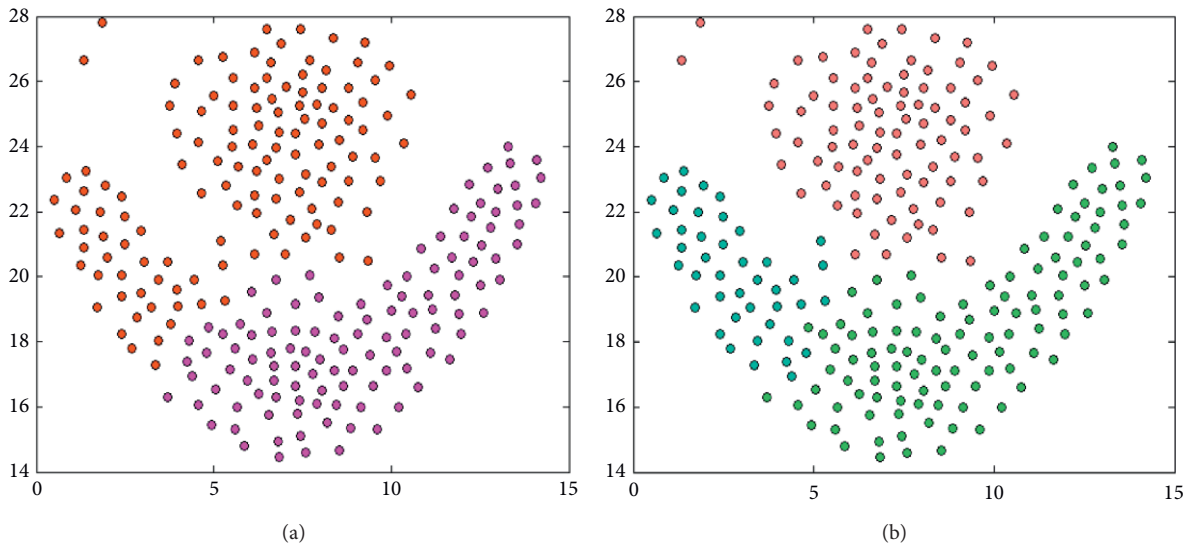


Figure 7: Continued.

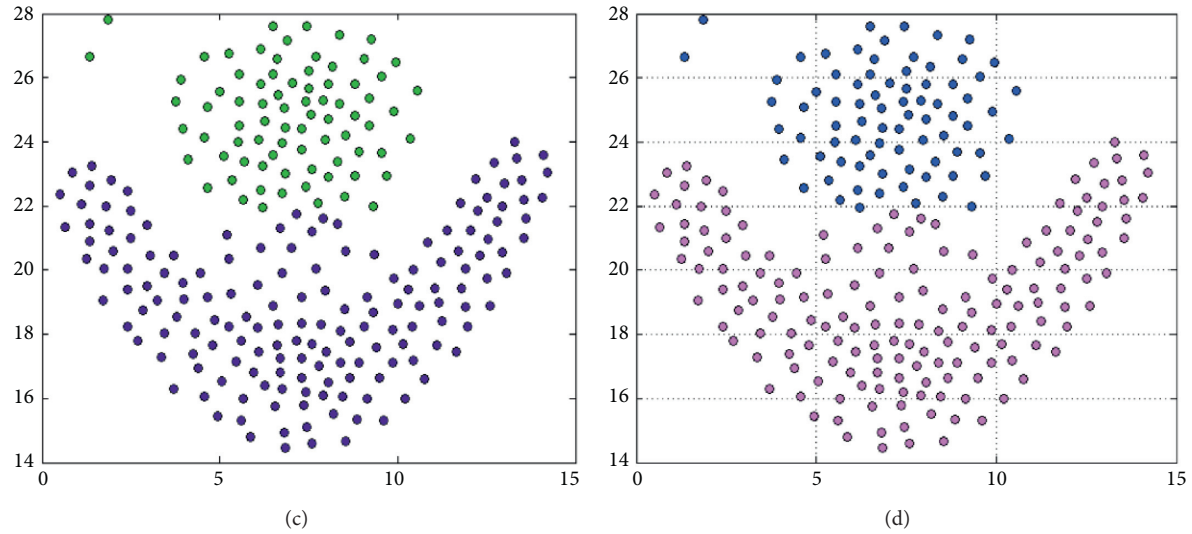


Figure 7: The clustering results of Flame data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

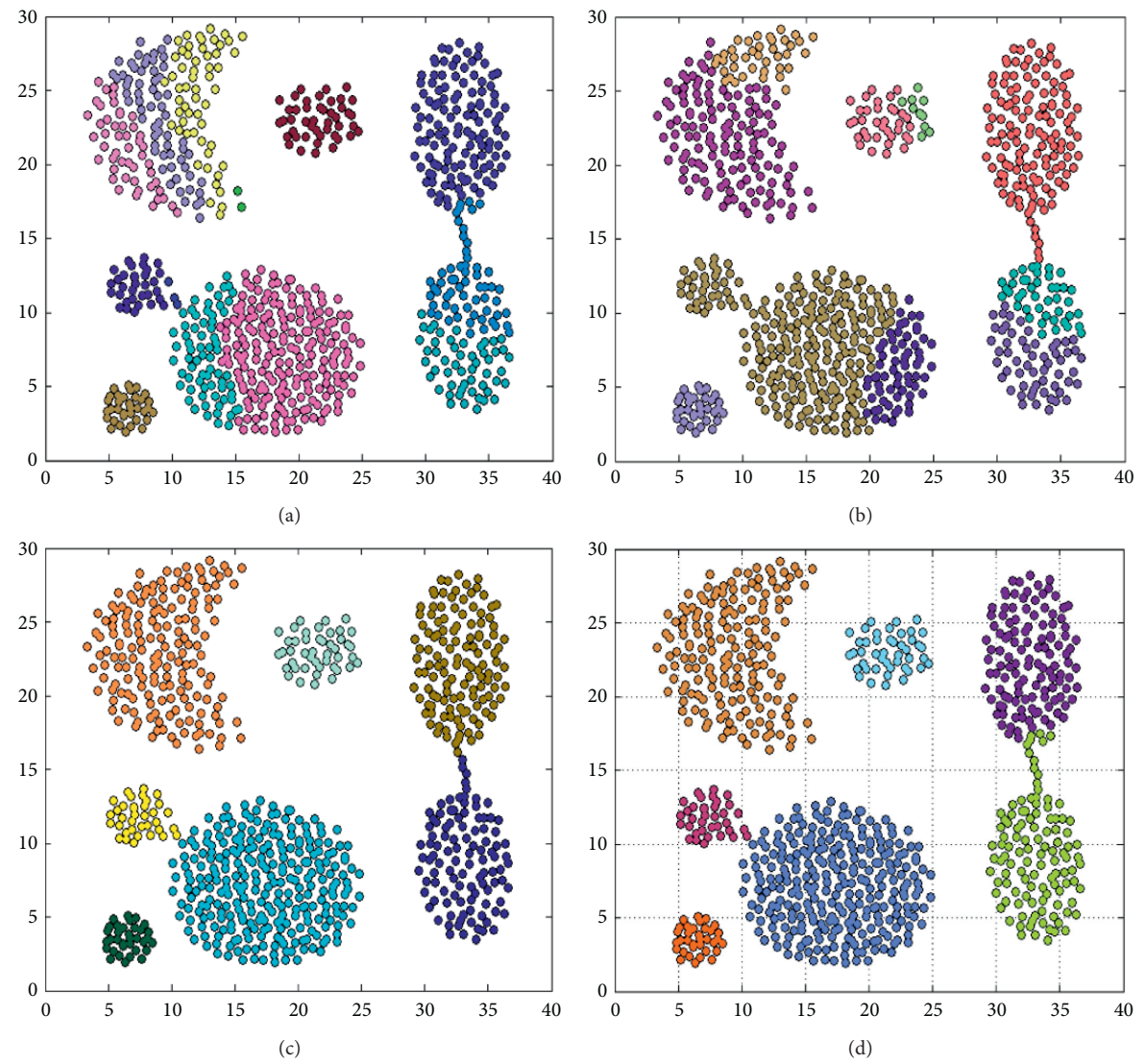


Figure 8: The clustering results of Aggregation data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

Table 4: The validity index value of the k-means.

Data set	K-means			
	FM	*F*1	Rand	Jaccard
D1	0.48	0.39	0.37	0.37
D2	0.45	0.41	0.33	0.28
Jain	0.57	0.49	0.41	0.39
Flame	0.29	0.27	0.19	0.19
Aggregation	0.35	0.33	0.24	0.21

Table 5: The validity index value of the DP.

Data set	DP			
	FM	*F*1	Rand	Jaccard
D1	0.81	0.79	0.71	0.68
D2	0.80	0.79	0.74	0.72
Jain	0.77	0.74	0.67	0.67
Flame	0.94	0.92	0.91	0.88
Aggregation	0.84	0.80	0.76	0.76

Table 6: The validity index value of the AP.

Data set	AP			
	FM	*F*1	Rand	Jaccard
D1	0.51	0.48	0.43	0.43
D2	0.49	0.48	0.41	0.39
Jain	0.55	0.53	0.47	0.41
Flame	0.54	0.50	0.44	0.83
Aggregation	0.61	0.69	0.81	0.88

Table 7: The validity index value of the DPKT-AP.

Data set	DPKT-AP			
	FM	*F*1	Rand	Jaccard
D1	0.79	0.76	0.76	0.71
D2	0.77	0.73	0.68	0.67
Jain	0.71	0.71	0.70	0.69
Flame	0.82	0.77	0.71	0.71
Aggregation	0.69	0.64	0.60	0.57

6. Conclusion

The outstanding contributions of this paper include combining the advantages of the DP algorithm and k-means algorithm with the original AP algorithm and proposing the improved integrated clustering learning strategy based on three-stage affinity propagation algorithm with density peak optimization theory (DPKT-AP). DPKT-AP has the advantage of high clustering accuracy. In view of that, the AP algorithm was suitable for processing spherical data, the DPKT-AP obtained the subgroups with spherical structures in advance by using the DP and k-means algorithms, and finally, the clustering process of the AP was carried out. Thus, better clustering results are obtained. Simulation results demonstrate that the DPKT-AP algorithm can reduce the difficulty of the clustering process for different size, structure, and density data sample and improve the clustering performance. Compared with the traditional algorithm, the proposed algorithm has obvious advantages. Of course, there are still some limitations in the proposed DPKT-AP, for example, higher time cost due to number of iterations and insufficient ability to identify outliers. In the future work, with regard to the situation that the clustering effect of high-dimensional data is weaker than the counterpart of lower-dimensional data and the remaining limitations, we will introduce a function which combines the density with distance or change the distance calculation method for the further study [32–37].

Data Availability

The data sets in the paper are available in http://cs.uef.fi/sipu/datasets/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Limin Wang and Wenjing Sun contributed equally to this work.

Acknowledgments

This study was supported by the National Science Foundation of China under Grant nos. 61572225 and 61472049, Foundation of Jilin Provincial Education Department under Grant no. JJKH20190724KJ, Jilin Province Science & Technology Department Foundation under Grant nos. 20190302071GX and 20200201164JC, and Development and Reform Commission Foundation of Jilin province under Grant no. 2019C053-11.

References

[1] V. Oona, "Using data mining methods to solve classification problems in financial-banking institutions," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 54, no. 1/2020, pp. 159–176, 2020.

[2] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 77–92, 2019.

[3] M. Miheala and C. Cristian, "Developing an index score for the internal auditor profile in Romania based on real data analysis," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 93–111, 2019.

[4] J. Yoon and S. Joung, "A big data based cosmetic recommendation algorithm," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 40–52, 2020.

[5] E. Graham, J. Heidelberg, and B. Tully, "BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation," *PeerJ*, vol. 5, 2017.

[6] L. Wang and S. Cheng, "Data-driven resource management for ultra-dense small cells: an affinity propagation clustering approach," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 3, pp. 267–279, 2018.

[7] S. Zhou and Z. Xu, "Automatic grayscale image segmentation based on affinity propagation clustering," *Pattern Analysis and Applications volume*, vol. 23, pp. 331–348, 2020.

[8] A. Aizpurua and W. Koutstaal, "A new index of semantic short-term memory: development and validation of the conceptual span task in Spanish," *Plos One*, vol. 13, no. 12, 2018.

[9] D. Chen, J. Sheng, J. Chen, and C. Wang, "Stability-based preference selection in affinity propagation," *Neural Computing and Applications*, vol. 25, no. 7-8, pp. 1809–1822, 2014.

[10] W. Zhang, X. Wu, W.-P. Zhu, and L. Yu, "Unsupervized image clustering with SIFT-based soft-matching affinity propagation," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 461–464, 2017.

[11] S. Ding, M. Du, T. Sun, X. Xu, and Y. Xue, "An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood," *Knowledge-Based Systems*, vol. 133, pp. 294–313, 2017.

[12] H. K. Aljobouri, H. A. Jaber, O. M. Koçak, O. Algin, and I. Çankaya, "Clustering fMRI data with a robust unsupervised learning algorithm for neuroscience data mining," *Journal of Neuroscience Methods*, vol. 299, pp. 45–54, 2018.

[13] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification," *Applied Intelligence*, vol. 49, no. 4, pp. 1245–1259, 2019.

[14] G. Liu, V. P. Andreev, M. E. Helmuth et al., "Symptom based clustering of men in the LURN observational cohort study," *Journal of Urology*, vol. 202, no. 6, pp. 1230–1239, 2019.

[15] C. Yang, S. Liu, L. Bruzzone, R. Guan, and P. Du, "A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1152–1156, 2013.

[16] L. Wang, Z. Hao, and W. Sun, "A novel self-adaptive affinity propagation clustering algorithm based on density peak theory and weighted similarity," *IEEE Access*, vol. 7, pp. 175106–175115, 2019.

[17] L. Wang, Q. Ji, and X. Han, "Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity," *Tehnicki Vjesnik-Technical Gazette*, vol. 23, no. 2, pp. 425–435, 2016.

[18] Y. Wang, W. Pang, Y. Zhou, R. Zhou, K. Zheng, and M. Liu, "Density propagation based adaptive multi-density clustering algorithm," *Plos One*, vol. 13, no. 7, 2018.

[19] L. Wang, M. Li, X. Han, and R. Zhou, "Improved density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster centre," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 2, pp. 536–545, 2018.

[20] L. Wang, Z. Hao, and X. Han, "Gravity theory-based affinity propagation clustering algorithm and its applications," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 4, pp. 1125–1135, 2018.

[21] Z. Geng, R. Zeng, Y. Han, Y. Zhong, and H. Fu, "Energy efficiency evaluation and energy saving based on DEA integrated affinity propagation clustering: case study of complex petrochemical industries," *Energy*, vol. 179, pp. 863–875, 2019.

[22] F. Xu, X. Shu, X. D. Zhang, and B. Fan, "Automatic diagnosis of microgrid networks' power device faults based on stacked denoising autoencoders and adaptive affinity propagation clustering," *Complexity*, vol. 2020, Article ID 8509142, , 2020.

[23] Z. Wei, Y. Wang, S. He, and J. Bao, "A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection," *Knowledge-Based Systems*, vol. 116, pp. 1–12, 2017.

[24] L. Wang, X. Zhou, Y. Xing, M. Yang, and C. Zhang, "Clustering ECG heartbeat using improved semi-supervised affinity propagation," *IET Software*, vol. 11, no. 5, pp. 207–213, 2017.

[25] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[26] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[27] Y. S. Park and J. H. Choi, "Algorithm of three-party combined judgment analysis engine for earthquake early warning system," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 45–64, 2019.

[28] J. B. Kim, "Implementation of artificial intelligence system and traditional system: a comparative study," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 135–146, 2019.

[29] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering approach based on mini batch kmeans for intrusion detection system over big data," *IEEE Access*, vol. 6, pp. 11897–11906, 2018.

[30] Y. Liu, Z. Ma, and F. Yu, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowledge-Based Systems*, vol. 133, pp. 208–220, 2017.

[31] B. Wu and B. M. Wilamowski, "A fast density and grid based clustering method for data with arbitrary shapes and noise," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1620–1628, 2017.

[32] L. Fu and Y. Dong, "Research on internet search data in China's social problems under the background of big data," *Journal of Logistics, Informatics and Service Science*, vol. 5, no. 2, pp. 55–67, 2018.

[33] C. Zhang, M. Ni, H. Yin, and K. Qiu, "Developed density peak clustering with support vector data description for access network intrusion detection," *IEEE Access*, vol. 6, pp. 46356–46362, 2018.

[34] X. Xu, S. Ding, and Z. Shi, "An improved density peaks clustering algorithm with fast finding cluster centers," *Knowledge-Based Systems*, vol. 158, pp. 65–74, 2018.

[35] M. Du, S. Ding, X. Xu, and Y. Xue, "Density peaks clustering using geodesic distances," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1335–1349, 2018.

[36] A. Sajjad and F. Mehdi, "Particle swarm optimization algorithm for the prepack optimization problem," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 289–307, 2018.

[37] J. Jiang, Y. Chen, D. Hao, and K. Li, "DPC-LG: density peaks clustering based on logistic distribution and gravitation," *Physica A: Statistical Mechanics and Its Applications*, vol. 514, pp. 25–35, 2019.

Hindawi
Complexity
Volume 2021, Article ID 6610767, 12 pages
https://doi.org/10.1155/2021/6610767

Research Article

Evolution Mechanism of Advanced Equipment Manufacturing Innovation Network Structure from the Perspective of Complex System

Jianbo Wang[1,2] **and Xing Cao**[1,3]

[1]*Business School, Central South University, Changsha 410083, China*
[2]*Hunan University of Humanities, Science and Technology, Loudi 417000, China*
[3]*Hunan First Normal University, Changsha 410205, China*

Correspondence should be addressed to Xing Cao; caoxingsxy418@csu.edu.cn

Received 23 October 2020; Revised 21 November 2020; Accepted 14 December 2020; Published 6 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

Our country's equipment manufacturing industry ranks among the best in all developing countries, but compared with developed countries, there is still a long way to go. It is not only the backwardness of various technologies, but also the interference of other countries. Although our country's equipment manufacturing industry is not as advanced as the advanced technology of developed countries, we still have to stick to our original aspirations, do not underestimate ourselves, and be good at absorbing and learning from the strengths of others to make up for our own weaknesses. While not working behind closed doors and while absorbing technology from other countries, we can make use of our strengths to make up for our weaknesses and develop our own industrial technology. This paper studies the evolution trend of innovation network structure and at the same time studies the evolution mechanism of advanced equipment manufacturing innovation network structure from the perspective of complex systems. The explained variable in this article is green total factor productivity. The variable adopts the Malmquist–Luenberger global super-efficiency index model. There are two main explanatory variables. One is the heterogeneity that affects the efficiency of industrial evolution, including factor heterogeneity, structural heterogeneity, and environmental heterogeneity, and the other is the interaction term of equipment manufacturing specialization agglomeration degree dummy variable multiplied by factor heterogeneity. The regional economic development level is added to the model as a control variable. In the selection of measurement indicators, the per capita GDP is used as the control variable. The experimental results show that each sample is tested in pairs, and the standard error level of the mean is 0.018, which is less than 0.05, indicating that the efficiency of the equipment manufacturing industry's economic correlation spatial network has a significant impact on the overall economic development level of the industry. The reduction in spur helps to increase economic output.

1. Introduction

In recent years, with the increasing development of network technology and the increasing rise of networked organizations, innovation networks formed by the cooperation of different innovation entities have become a new organizational form for enterprise technological innovation activities. The formation of the innovation network is conducive to the accumulation of innovation resources, the improvement of innovation efficiency, and the transformation of scientific and technological achievements. An effective innovation model that can adapt to the current fierce competition only needs to be based on the innovation resources in the park or region, and in terms of time and quantity, it will be far from meeting the innovation and development needs of the cluster system. It is also necessary to adjust the innovation behavior of industrial parks accordingly, that is, from relying on technology learning and knowledge dissemination among local industrial park members, to actively developing external network

connections to obtain the required external high-level knowledge resource cluster innovation. Initially, static innovation activities that focused only on acquiring one-time knowledge and technical learning became dynamic innovation activities and finally formed continuous innovation through a series of stages and processes.

The world economy has entered a period of relatively low growth. The economies of large developed countries have fallen into recession, and the economic growth of developing countries has slowed. Developed countries such as European countries and the United States have increased their support for the real economy and implemented the "industrialization" strategy in order to seek in-depth adjustments to the economic structure. New changes have taken place in global economic growth. With the growth of market demand for individualization and adaptability, trade protectionism in some developed countries has increased, and some developing countries have also actively participated in the global industrial division of labor due to their advantages in work and raw materials. Besides, other agents donated, accelerated the transfer of industry and capital, and opened up the international market space. Guan gave a model of the evolution of the innovation network structure from a quantitative perspective and pointed out that the structural evolution of the cluster innovation network is mainly reflected in the addition, edge addition, and reconnection behavior of the cluster innovation network [1]. Piccinno et al. used complex network theory to give three simulation models for the evolution of the innovation network structure and calculated the structural characteristics of the innovation network under the three models with the simulation of technology, capital, and market nodes [2]. Li studied the formation and evolution mechanism of cluster innovation networks for specific industrial clusters [3].

China's economic growth has slowed from high-speed growth to medium-high growth, and downward pressure on the national economy has increased. Improving the quality and efficiency of growth has become the main goal of future economic growth. Some contradictions and problems concealed by the period of rapid economic growth are concentrated, forcing enterprises to accelerate their transformation and upgrading. The traditional factor-based and investment-based models no longer meet the requirements of the new economic normal, and the structural reform of the supply side has become a new trend. In the new economic situation, speeding up the transformation of the economic structure and industrial upgrading must have the rise and support of a number of emerging industries characterized by high technology. Hertem et al. conducted an empirical study on the innovation network of the equipment manufacturing industry cluster in Shenyang, but it mainly focused on the node elements and connection mode elements of the equipment manufacturing industry cluster innovation network [4]. Karlik and Platonov analyzed the composition of the equipment manufacturing cluster innovation network from a static perspective and described the innovation network [5]. Zhong summarized and analyzed the structural characteristics of the cluster innovation network, took the typical three types of industrial clusters in the west as an example, and studied the structural characteristics of the cluster innovation network [6].

There is no doubt that China is a large producer and a global factory, and it has always provided a large number of industrial products to the world. However, this situation is in exchange for a large amount of cheap labor in our country. The current growth model has reached its limit, and other ways must be found to rescue the manufacturing industry. The novelty of this article lies in the application of entropy method to the innovation and development evaluation of equipment manufacturing industry. Through the research and analysis of the development status of the equipment manufacturing industry and the factors affecting its innovation and development, an evaluation index system for the innovation development of the equipment manufacturing industry has been established, and an empirical analysis of the innovation development of the equipment manufacturing industry has been carried out.

2. Evolution Mechanism of Advanced Equipment Manufacturing Innovation Network Structure from the Perspective of Complex System

2.1. Connotation and Characteristics of Equipment Manufacturing Cluster. The equipment manufacturing cluster is different from the general cluster and has its own uniqueness. However, from the current domestic research, there is currently no unified concept. Different scholars have given their own opinions from different angles [7]. In order to further analyze the composition of the equipment manufacturing cluster and its innovation network, it will be defined from the perspective of dynamic evolution [8].

2.1.1. Connotation and Characteristics of Equipment Manufacturing Industry. According to the definition of relevant domestic research institutions and scholars, the equipment manufacturing industry is also called the equipment industry [9]. It is a general term for industries that manufacture various technical equipment to meet the development of various sectors of the national economy and national security needs. According to the classification of the national financial industry, we can divide the equipment into five categories according to the content of knowledge and the technical difficulty of the product, supplemented by the needs of direct control by the country: general equipment, basic equipment, basic high-tech equipment, complete equipment, and safety equipment [10]. Compared to the light processing industry, the main features of the equipment manufacturing industry are complex technical composition, a variety of supporting elements, large companies with strong technological capabilities of integration and innovation, and higher capital and labor requirements. With the process of global integration, global cooperation in technology research, development, production, and sales is increasingly strengthened.

2.1.2. Composition of Equipment Manufacturing Cluster Innovation Network. Cluster innovation is a process in which the enterprises in the cluster continuously acquire knowledge from the formal and informal network of relationships in which they are located, and integrate them [11]. The main bodies of the equipment manufacturing cluster have formed various close relationships that contribute to the improvement of innovation capabilities due to their exchanges and connections with each other, and they have obvious network characteristics [12]. The equipment manufacturing cluster innovation network is as follows: under a specific regional social and cultural background, enterprises and related institutions in the equipment manufacturing cluster establish a long-term and relatively stable form that can promote innovation within the cluster through communication with each other.

2.1.3. Node Elements of Equipment Manufacturing Cluster Innovation Network. The node elements of the equipment manufacturing cluster innovation network mainly include enterprises, universities/scientific research institutions, intermediary service organizations, and governments [4]. Different from the general cluster innovation network, the center and main body of the equipment manufacturing cluster innovation network are the leading enterprises in the cluster. A large number of small- and medium-sized supporting enterprises have gathered around these leading enterprises. Due to the high technological content of the equipment manufacturing industry, although some large leading companies have mastered a large number of leading technological advantages through introduction, digestion, and absorption, they must have their own proprietary knowledge to obtain sustained competitive advantages [13]. To this end, some large leading companies have also established long-term cooperative relationships with universities, scientific research institutes, and vocational and technical schools within and outside the cluster to expand technological research and development capabilities and enhance independent innovation capabilities. The equipment manufacturing industry is a highly capital-intensive and policy-oriented industry, so the government plays an important role in the cluster innovation network. On the one hand, it can help enterprises in the cluster to win more national, provincial, and ministerial projects; on the other hand, it can provide various policy support for cluster enterprises to improve the level of cooperation and social cooperation between cluster entities in the region. The capital stock ensures the efficient transfer of knowledge and information within the cluster.

2.1.4. Relational Connection of Equipment Manufacturing Cluster Innovation Network. Similar to the general cluster innovation network, the relationship connection between the subjects includes both formal cooperation agreements and informal exchanges and communication, and with the number of relationships between the subjects, the connection can be strong or weak [14]. Since the difference between equipment manufacturing clusters and general industrial clusters is mainly reflected in the enterprise nodes, this section focuses on the relationship between enterprises and enterprises. In addition to the relational connection of the industrial chain model, there is also a strong competition-cooperative model relational connection between cluster enterprises. Due to the differences in enterprise scale and capabilities, a group of flying geese is usually formed with large enterprises as the core, medium-sized enterprises as the second echelon, and small enterprises as auxiliary. The competition among cluster enterprises is mainly manifested as the grabbing of market and resources among enterprises at the same level [15]. Generally speaking, the number of leading companies in a cluster is small, and there is no strong competitive relationship between them, which is mainly reflected in the plundering of external markets and resources with other companies of the same type outside the cluster. In order to reduce costs and gain a competitive advantage, they have formed a close supporting cooperative relationship with small- and medium-sized supporting enterprises, namely, cooperative relations.

2.2. Promotion Mechanism of Technological Progress on Industrial Upgrading. Industrial upgrading is a process of continuous development and change based on technological progress. From the perspective of supply, technological advancement promotes industrial upgrading by creating new industries, providing new products, changing factor supply ratios, promoting industrial integration, and improving the quality of human capital, thereby promoting industrial upgrading and thus continuing to increase productivity [16]. The technological structure upgrade and the industrial structure upgrade show a one-to-one correspondence in a relatively long period of time.

2.2.1. Technological Progress Promotes Industrial Upgrading by Creating Emerging Industries and Eliminating Backward Industries. Each product must go through the start-up phase, the growth phase, the maturity phase, and the decline phase and finally be replaced by a newer product. The impact of changes in demand structure on the industrial structure is immediate. Technological progress is an important factor in promoting product upgrades and creating new industries [17]. There are two emerging industries in the field of equipment manufacturing: One is an emerging industry, which refers to an industry that has gradually formed due to technological discovery. In an industry that has never been seen before, major technological advances will stimulate new demand for the formation of a new industry [18]. The second is industrial integration, which refers to the formation of new industries by integrating different subindustries within an industry or between different industries on the basis of continuous technological integration. Emerging industries created through technological integration can usually meet market demand and occupy a larger market share.

2.2.2. Technological Advancement Promotes Industrial Upgrading by Providing New Tools, New Processes, and New Methods. The three elements of technology are tools, methods, and processes. Technological progress has

promoted the emergence of new tools, new methods, and new processes; improved the use efficiency of factor inputs; reduced the consumption of energy and resources in the production process; and thus reduced costs. Due to the emergence of new tools, new methods, and new processes, the traditional equipment manufacturing industry can also undergo technological transformation, and its product quality will be improved accordingly [19]. At the same time, affected by this, upstream and downstream enterprises in the industrial chain will also carry out technological upgrades to adapt to the coordinated development of the industrial chain.

2.2.3. Technological Progress Promotes Industrial Upgrading by Changing the Proportion of Factor Input Allocation. The rate input for each subindustry is different from the ratio of the equipment industry. Because each industry has different technical levels and technical capabilities, and the speed and effect of technological progress are also different, the elasticity of demand in each industry will change, leading to demand. The rate input for each subindustry is different from that of the equipment industry. Because each industry has different technical levels and technical capabilities, and the speed and effect of technological progress are also different, the demand elasticity of each industry will change, leading to changes in the demand structure [20]. When technological progress and other reasons increase the market demand for new products, the profitability of capital will push a large number of producers into the new product industry, thereby promoting the internal upgrading of the industry.

2.3. Model Description. In the classic BA model, the premise of its assumption is that there are isolated points in the cluster innovation network. However, from the actual network connection, in addition to formal relationship links, there are also informal exchanges and communication, and there are many types of nodes in the network, so the probability of outliers in the network is very small [21].

On the premise of satisfying complementary resources, any individual first selects the "local world" of the node newly added to the network when contacting other individuals in the network and then contacts other members of the "local world" network. This is the so-called "cluster network", and the probability of preferential connection is

$$p(k_i) = \frac{k_i}{\sum_{j\in N} k_j}. \tag{1}$$

N represents the node set formed by all nodes, k_i represents the node degree of node i, and k_j represents the node degree of node j.

Add m new edges to the network with probability $p\,(0 \le p < 1)$. In the initial network, a node is randomly selected as the starting point, and the other end is selected according to (1). Repeat the process $m\,(0 \le m < m_0)$ times. At this time, the degree value of node i is

$$\left(\frac{\partial k_i}{\partial t}\right) = pm\frac{1}{N} + pm\frac{k_i}{\sum_{j\in N} k_j}. \tag{2}$$

Add m edges to the network with probability $q\,(0 \le p < 1)$. First, randomly select an edge of a node. After disconnecting this edge, connect to node j, and follow the probability of (1) when selecting j. Repeat this process m times. At this time, the degree value of the node i is

$$\left(\frac{\partial k_i}{\partial t}\right) = -qm\frac{1}{N} + qm\frac{k_i}{\sum_{j\in N} k_j}. \tag{3}$$

Add a new node with probability $1 - p - q\,(0 \le p + q < 1)$. The new node is connected with the existing m nodes in the graph according to the probability of (1). At this time, the degree value of node i is

$$\left(\frac{\partial k_i}{\partial t}\right) = (1 - p - q)m\frac{k_i}{\sum_{j\in N} k_j}. \tag{4}$$

From comprehensive formula (2)–(4), we have

$$\frac{\partial k_i}{\partial t} = (p - q)m\frac{1}{N} + m\frac{k_i}{\sum_{j\in N} k_j}, \tag{5}$$

where p represents the probability of adding edges, q represents the probability of reconnection, and m represents the number of nodes or edges added to the center of the network. N represents the number of nodes in the entire network. As time changes, it will gradually meet the following conditions:

$$\begin{aligned} N_1 &= m_0 + (1 - p - q)t, \\ \sum_j k_j &= mt\,(1 - p - q) + \sum_j k_j\,(0). \end{aligned} \tag{6}$$

When t is large enough, the influence of m_0 and $\sum_j k_j\,(0)$ on N_t and $\sum_j k_j$ is small and can be ignored:

$$N_t = (1 - p - q)t, \tag{7}$$

$$\sum_j k_j = mt\,(1 - p - q). \tag{8}$$

From the combination of (4) and (8), we can get

$$\frac{\partial k_i}{\partial t} = \frac{1}{t}\left(b + \frac{k_i}{1 + p - q}\right). \tag{9}$$

And because of $(\mathrm{d}k_i/(b + (k_i/1 + p - q))) = (\mathrm{d}t/t)$, we can get

$$k_i(t_i) = C(1 + p - q)t^{(1/(1+p-q))} - b(1 + p - q). \tag{10}$$

Let $k_i(t_i) = m$; then,

$$C = \frac{m + b(1 + p - q)}{(1 + p - q)bt_i^{(1/(1+p-q))}},$$

$$k_i(t) = (m + b(1 + p - q))\left(\frac{t}{t_i}\right)^{(1/(1+p-q))} - (1 + p - q)b. \tag{11}$$

Among them,

$$b = \frac{p-q}{1-p-q}m. \tag{12}$$

In order to concretely describe the evolution of the equipment manufacturing industry cluster innovation network structure, below we will combine the phase characteristics of this evolution.

2.4. Characteristics of Technological Innovation Network of High-Tech Enterprises

2.4.1. Highly Complementary Knowledge Resources among Innovative Subjects. The problem of the optimal use of distributed knowledge is certainly not a new problem, it is only caused by the arrival of the knowledge economy, and any complex social system will face it. With the advent of the knowledge economy, the intensity of product knowledge has increased, and there is an increasing need to combine different distributed knowledge from different sources. In the high-tech industry, product knowledge is distributed in the hands of different production companies in the industry, and everyone continues to deepen their knowledge in a specific field. However, when the whole product is produced, it can be integrated with each other. Therefore, the high complementarity of innovation resources is very obvious in the high-tech industry [22]. From another perspective, when high-tech companies choose cooperative innovation partners, they generally examine the two important factors influencing cooperation complementarity and technological leadership. The complementarity of cooperation includes two aspects: one is the complementarity of technical resources, and the other is the complementarity of organization and management. Generally speaking, the more complementary the technical resources of the parties involved in the technological innovation network, the greater the performance produced by the technological innovation network.

2.4.2. Uncertainty in the Development of Technological Innovation Networks. High technology has the characteristics of highly intensive knowledge, so high technology has strong crossover and integration. In the development process, with the help of the original accumulation and innovation of technology, further development and innovation will form a qualitative leap in science and technology under new historical conditions. It is precisely because high technology is at the forefront of science and technology, so any pioneering ideas, design, and implementation methods are uncertain, and many research results are difficult to predict. If the gradual model represents the evolution of quantitative changes, then the transition model means the evolution of qualitative changes. Although the form of the product is inextricably linked with the past, the basic technology is far away [23]. Technological changes are usually caused by two approaches to solve real-world problems or capture new technology and market opportunities. Technological changes usually lead to transitions in technological orbits, and changes in technological orbits usually require new core capabilities. This shows from another perspective that this change in technology will increase the complementary resources in the network, which will lead to the expansion of the technological innovation network to obtain these complementary resources. Corresponding to these processes, the learning of technological innovation network will change, from the previously narrow knowledge field to a broader knowledge and technology field.

2.4.3. Overall Importance of Social Capital. In the embedded network relationship, companies with a large amount of social capital have many information channels, which can attract favorable partners and alliances and negotiate favorable conditions for themselves. The integrated relationship increases the knowledge exchange between companies, which is the basis for maintaining the partnership. Interinstitutional trust is a prerequisite and an important stimulus for knowledge exchange and innovation. This process requires the support of social capital from high-tech companies. Collective learning transports knowledge to space, which is a dynamic process of free knowledge transfer between entities [24]. Collective learning is a process of knowledge flow caused by the interaction between organizations in a network of technological innovation and a social process of knowledge accumulation. This process is based on the rules and procedures followed by all members of the network, which can go beyond individual rationality and draw on collective logic. The main body of network innovation responds to the uncertainties faced by the technological innovation network through collective observation, evaluation, and coherent measures (Figure 1).

3. Evolution Mechanism of Advanced Equipment Manufacturing Innovation Network Structure from the Perspective of Complex Systems

3.1. Demand Analysis. The "independent" upgrade path of system engineering of the model aims not only to increase investment in independent innovation, but more importantly to accurately answer the specific methods of independent innovation to promote industrial upgrading. With the advancement of technology, the enabling role of information technology has been further demonstrated, and it is transforming from information technology supporting business development to information technology leading innovation-driven and transformational development. The theory of technological innovation and industrial upgrading takes the technological accumulation formed by the continuous development of the enterprise as the main driving force and is based on the realization of regional expansion. Due to the improvement of the level of technological development, overseas investment models have gradually shifted from resource acquisition to technology acquisition. In addition, related industries for overseas investment are gradually upgrading, and their composition is related to the

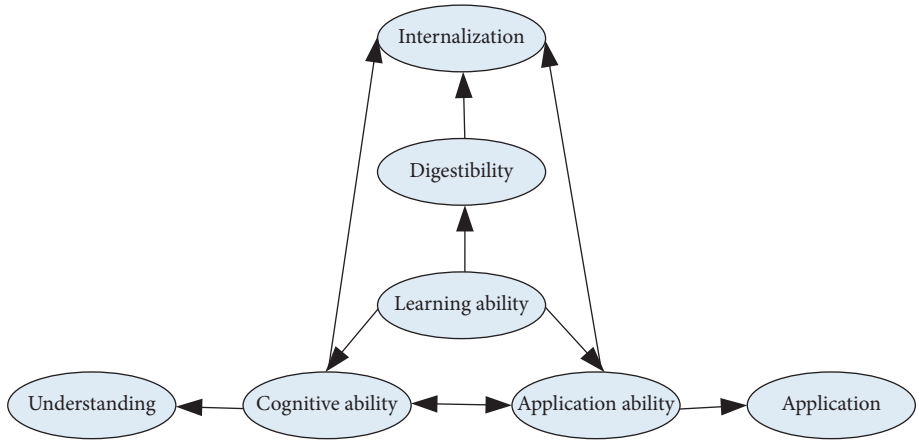


Figure 1: Relation diagram of equipment manufacturing cluster innovation network.

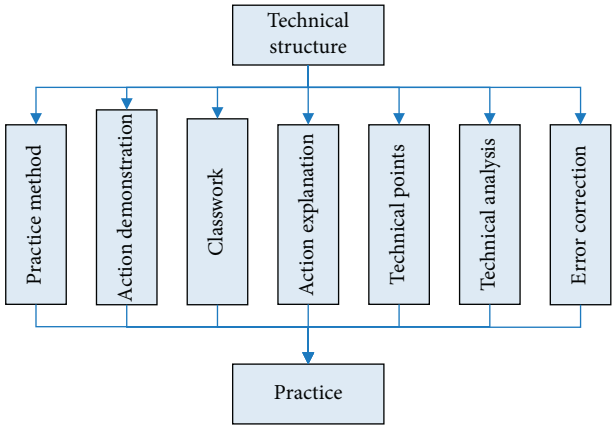


Figure 2: The structural mode of the technical content of gymnastics courseware.

adjustment of regional location distribution. It is of great significance for emerging countries to use overseas investment to achieve technological progress and upgrades, thereby further optimizing the industrial structure and increasing the level of international competition. The overall framework of system requirements is shown in Figure 2.

3.2. Test Subject. The variables studied in this experiment mainly include three categories: explained variables, main explanatory variables, and control variables. The explained variable in this paper is green total factor productivity, and the variable uses the Malmquist–Luenberger global super-efficiency index model. There are two main explanatory variables. One is the heterogeneity that affects the efficiency of industrial evolution, including factor heterogeneity, structural heterogeneity, and environmental heterogeneity, and the other is the virtual variable and factor of the equipment manufacturing industry's specialization concentration. The interaction terms of heterogeneity multiply, the level of regional economic development is added to the model as the control variable, and the per capita GDP is used as the control variable in the selection of measurement indicators. The improvement of green total factor productivity is closely related to the level of regional economic development. Generally speaking, a high level of regional economic development will bring economic support to the green development of the equipment manufacturing industry, such as sound infrastructure and a large amount of consumption, demand, etc. In addition, the ability to introduce or develop advanced eco-environmental protection technologies and carry out larger-scale product and service innovations is conducive to promoting the rapid growth of green total factor productivity.

3.3. Model Building. The integration of vehicle control skills has two situations, promoting effect and hindering effect on the total factor productivity of China's equipment manufacturing industry. Considering the nonlinear relationship between vehicle control skill integration index and total factor productivity, this paper introduces the square term of integration index to analyze the nonlinear curve relationship under the influence of the two effects. This paper adds industry variables such as R&D investment intensity, property rights characteristic factors, export dependence, condition construction intensity, number of high-quality scientific and technological personnel, and time variables as control variables to control the influence of industry characteristics and time trends. In order to enable one-to-one correspondence of data, this paper merges the general equipment manufacturing industry and special equipment manufacturing industry into general and special equipment manufacturing industries. Therefore, all control variables are merged into 6 industries. The classification standards are the same.

3.4. Data Processing. The index selected in this article is the annual statistical data of books, which is the index data of practical value. Therefore, the trapezoidal fuzzy table distribution function is usually used to calculate the participation of each indicator. The specific formula is as follows:

Large trapezoidal distribution:

$$r(x) = \frac{x - c}{d - c}, \quad c < x < d. \tag{13}$$

Small trapezoidal distribution:

$$r(x) = \frac{b - x}{x - a}, \quad a < x < b. \tag{14}$$

For the ideal score, it can usually be set to the middle value of each interval. This paper divides the evaluation grades into five grades, and the index data after normalization processing all fall into the [0-1] interval.

There are many data standard processing methods, but different data standardization methods will have a certain impact on the evaluation results of the system. For the positive indicator standardization method,

$$y_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}}. \tag{15}$$

For the negative index standardization method,

$$y_{ij} = \frac{\max\{x_{ij}\} - x_{ij}}{\max\{x_{ij}\} - \min\{x_{ij}\}}. \tag{16}$$

After standardizing the data, using the principal component analysis of nonlinear logarithmic centering, the processing steps of logarithmic transformation and row vector centering are

$$z_{ij} = \ln y_{ij} - \sum_{i=1}^{m} \frac{\ln y_{ij}}{m}. \tag{17}$$

3.5. Statistical Methods. SPSS 23.0 software was used for data processing, and the count data was expressed in percentage (%); k is the amount of data in this experiment, σ^2 is the variance of all survey results, and $P < 0.05$ indicates that the difference is statistically significant. The formula for calculating reliability is shown in

$$a = \frac{k}{k - 1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right). \tag{18}$$

4. Evolution Mechanism of Advanced Equipment Manufacturing Innovation Network Structure from the Perspective of Complex System

4.1. Evaluation Index System Based on Index Reliability Testing. Reliability refers to the stability and consistency of the questionnaire. This article adopts the α coefficient method created by Cronbach. The α coefficient can be obtained by Reliability Analysis in SPSS software. It is generally believed that the α coefficient above 0.8 indicates that the effect of the index setting is very good, and above 0.7 is also acceptable. The results are shown in Table 1.

It can be seen from Table 1 that the cluster analysis of the four types of data of completed value added, sales revenue, profits and taxes, and profits in this experiment is acceptable ($\alpha > 0.7$), and there is no absolute difference between the various development stages. The pros and cons are not that the more complex the model, the better the prediction performance. Within the acceptable range, the preconditions for the experiment are met, which provides a basis for subsequent experimental analysis.

4.2. Evolution Trend of Innovation Network Structure. The visualization diagram and small-world measurement of the complex network simulation of the innovation network structure can only have a preliminary grasp and understanding of the network structure form as a whole. In order to fundamentally recognize and understand the evolution trend of the innovation network structure, this article investigates the network evolution trend from two specific indicators of network density and network relevance.

4.2.1. Evolution Trend of Innovation Network Density. The economic system and the mode of economic growth have established fundamental changes of great significance: one is the shift from a planned economy to a market economy, and the other is the shift from an extensive economy to an intensive economy. The two fundamental changes have greatly promoted the efficiency of the allocation of economic factors among regions and strengthened the links between economic entities. The evolution trend is shown in Figure 3.

It can be seen from Figure 3 that although a higher network density is conducive to strengthening the connection between innovative network structures, a higher density also means more connections, which also causes "redundant lines" in the network. If there are too many redundant connections, this will extend the connection path, increase the transaction cost, reduce the efficiency of resource and element allocation, and inhibit the speed of economic development. Therefore, it is necessary to consider maintaining a suitable network density to ensure the flow speed and configuration efficiency of the elements as much as possible.

4.2.2. Evolution Trend of Innovation Network Relevance. The network correlation characteristics of the comprehensive economic space of the urban agglomeration in the Yangtze River Delta are measured by using the network correlation degree, rank degree, and efficiency in the complex network analysis method. The result is shown in Figure 4.

It can be seen from Figure 4 that the continuous improvement of the market-oriented system and the continuous improvement of the supply and demand system and communication interaction of the factor market have promoted the market's fundamental role in resource allocation and have reduced the transaction costs of the flow and linkage of factors between regions. As a

Table 1: Summary table of reliability test results.

Intelligent analysis	Type of data	Alpha coefficient(α)
The embryonic stage of innovation network	Complete added value Sales revenue Profit and tax Profit	0.8636
The growth period of innovation network	Complete added value Sales revenue Profit and tax Profit	0.7742
The mature period of innovation network	Complete added value Sales revenue Profit and tax Profit	0.7384
Innovation network decline or change period	Complete added value Sales revenue Profit and tax Profit	0.7429

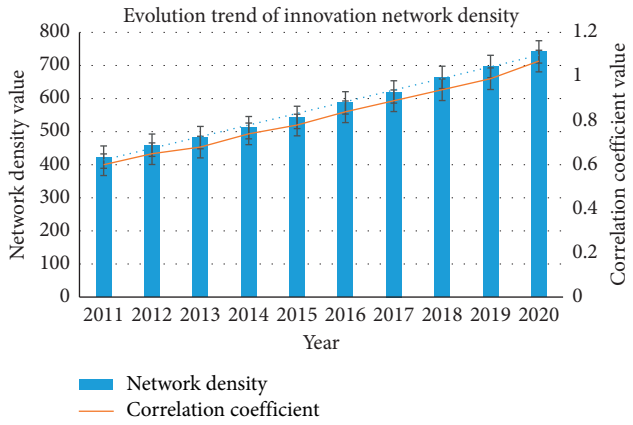


Figure 3: Analysis of the evolution trend of innovation network density.

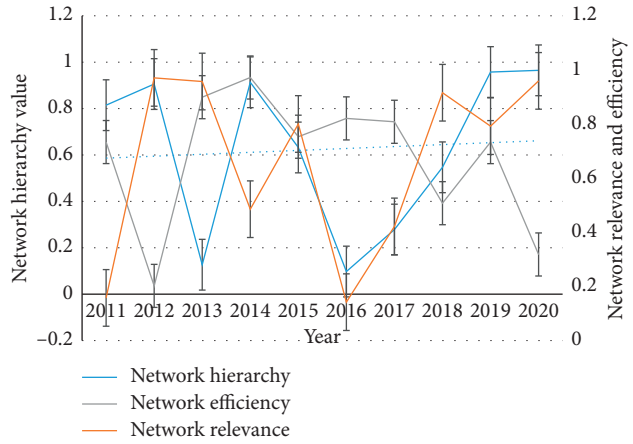


Figure 4: Analysis of the evolution trend of innovation network relevance.

result, the relationship between the economic space of the innovation network structure is increasing, and the stability of the innovation network structure is strengthened.

4.3. Experimental Evolution of Equipment Manufacturing Cluster Innovation Network Structure

4.3.1. Development of Equipment Manufacturing Industry Cluster. Through the development in recent years, the scale of small- and medium-sized supporting enterprises in the equipment manufacturing industry in the city has gradually expanded, the industrial concentration and specialization level have been greatly improved, and the scale benefits of industrial clusters have played a role. With the support of all walks of life, the equipment manufacturing industry strives to adjust the industrial structure and improve the quality of operation. The results are shown in Table 2. At the same time, the mean, standard deviation, and standard error of the mean are calculated, as shown in Figure 5.

It can be seen from Figure 5 that at the same time each sample is tested in pairs, and the standard error level of the mean is 0.14, which is greater than 0.05, indicating that the network density of the comprehensive economic correlation spatial network of the equipment manufacturing industry has a significant impact on the overall economic development level of the industry. The increase in network density can significantly increase economic output. It can be seen from Table 2 that, from the perspective of increasing the density of the network, in order to achieve an increase in economic output and rapid economic development, we should continue to attach importance and give full play to the fundamental role of the market in the allocation of resources and factors, and promote the spatial flow of factors.

4.3.2. Evolution of Equipment Manufacturing Innovation Network Structure Based on Logistic Algorithm. By adopting single-factor and multifactor control variable method, the equipment manufacturing innovation network structure model based on binomial logistic regression algorithm is established. In the process of parameter tuning, it mainly adjusts the maximum number of iterations of the model, the learning rate, and the maximum depth of the tree.

Table 2: Equipment manufacturing industry cluster development.

Year	Complete added value	Sales revenue	Profit and tax	Profit
2015	400.6	249.81	29.02	40.42
2016	460.3	324.78	21.94	60.1
2017	—	390.78	33.01	80.34
2018	570.9	—	121.2	—
2019	630.8	—	93.4	—
2020	700.5	—	—	—

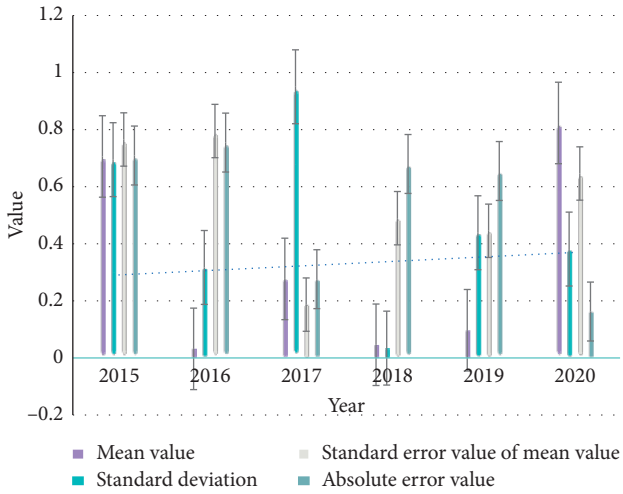


Figure 5: Analysis chart of equipment manufacturing cluster development.

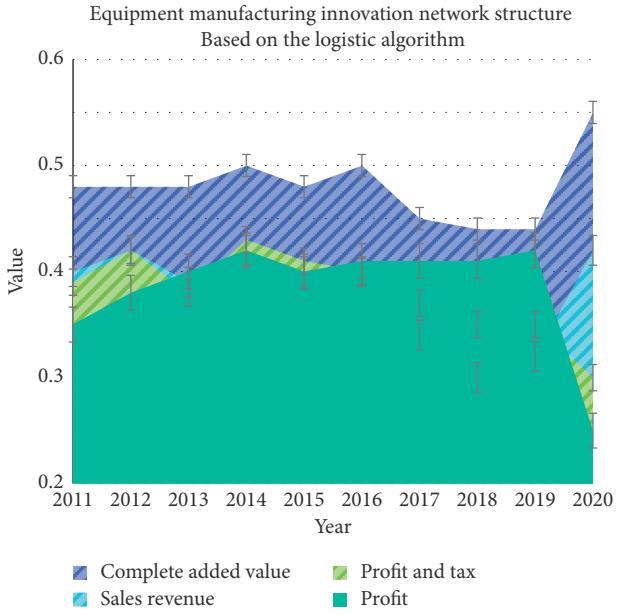


Figure 6: Analysis diagram of the evolution of the equipment manufacturing innovation network structure based on the logistic algorithm.

Each iteration will produce a weak learner. If the number of weak learners is too small, it is easy to underfit, and if there are too many, it is easy to overfit. The result is shown in Figure 6. At the same time, each sample is tested on the selected samples, and data such as the mean, standard deviation, and standard error of the mean are obtained, as shown in Table 3.

It can be seen from Table 3 that at the same time each sample is tested in pairs, and the standard error value of the mean is 0.004, which is less than 0.05. The network level of the economic correlation spatial network of the equipment manufacturing industry has a significant impact on the overall economic development level of the industry. The reduction of network level will help increase the economic output of the equipment manufacturing industry. It can be seen from Figure 6 that the decline in the level of the network means that the level of the network's hierarchical structure becomes smaller; that is, the subordinate status of the equipment manufacturing industry's economy has changed from the subordinate status of the one-way connection in the past to the equal status of the two-way connection.

4.3.3. Evolution of Equipment Manufacturing Innovation Network Structure Based on MMD Algorithm. By adopting single-factor and multifactor control variable methods, the equipment manufacturing innovation network structure model based on MMD algorithm is established. In the process of parameter tuning, it mainly adjusts the maximum number of iterations of the model, the learning rate, and the maximum depth of the tree. Each iteration will produce a weak learner. If the number of weak learners is too small, it is easy to underfit, and if there are too many, it is easy to overfit. The result is shown in Figure 7. At the same time, each

Table 3: Analysis table of the results of the risk assessment system of Yebes algorithm.

	Mean	Standard deviation	Standard error of the mean
Complete added value	2.32	9.462	1.526
Sales revenue	1.94	8.347	1.524
Profit and tax	−1.46	9.432	1.529
Profit	−1.21	8.462	1.527

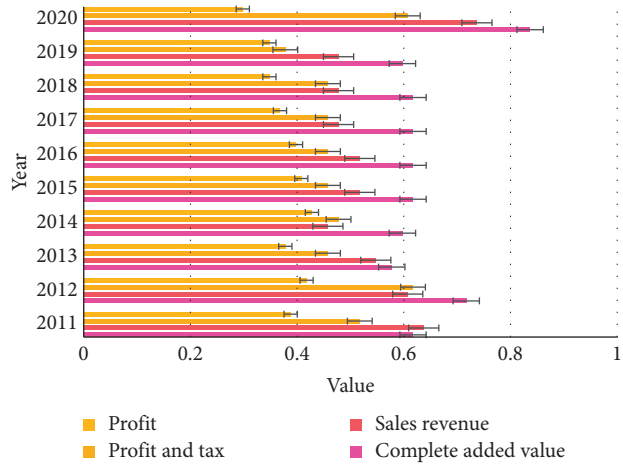


Figure 7: Research diagram of the evolution of the equipment manufacturing innovation network structure based on the MMD algorithm.

Table 4: Analysis table of the results of MMD algorithm risk assessment system.

	Mean	Standard deviation	Standard error of the mean
Complete added value	1.49	11.624	1.824
Sales revenue	1.42	11.523	1829
Profit and tax	−1.33	9.643	1.276
Profit	−1.01	9.427	1.271

sample is tested on the selected samples, and data such as the mean, standard deviation, and standard error of the mean are obtained, as shown in Table 4.

It can be seen from Table 4 that at the same time each sample is tested in pairs, and the standard error level of the mean is 0.018, which is less than 0.05, indicating that the network efficiency of the equipment manufacturing industry's economic correlation spatial network has a significant impact on the overall economic development level of the industry. The reduction of network efficiency helps to promote the increase of economic output. It can be seen from Figure 7 that we will continue to accelerate the integrated construction of the equipment manufacturing industry, further give full play to the decisive role of the market in resource allocation, strengthen the supply and demand links between parts, reduce the cost of economic space interaction, and gradually improve the stability of the spatial connection network.

5. Conclusions

From the perspective of network structure, China's equipment industry innovation network presents an example of a "core-periphery" structure. There are mainly scattered parallel structures, star-shaped dominant structures, and cooperative composite structures. Most of them do not have a coordinated complex structure. This shows that the current innovation network structure of China's equipment industry is unified and lacks stability. Establishing a multilevel collaborative innovation system and strengthening the innovation network of large companies have become necessary means to promote the development of the equipment industry. From the perspective of network characteristics, the current innovation network of China's equipment industry has low density, poor coherence, lack of central organization, insufficient overall network control, and low knowledge transfer efficiency. State-owned enterprises are the core of the network and continue to be the most important part of the equipment industry. The deficiencies of state-owned enterprises have also led to lagging innovation; private companies and consortia have limited flexibility and innovation capabilities, while universities and research institutions are at the edge of the network, and their contribution to innovation is low. Optimizing the industrial structure and exerting the vitality of different components constitute the way to future development.

Large infrastructure equipment companies must act as technology portals and importers of industrial complex

knowledge; establish external network relationships through active strategic activities, thereby introducing advanced external knowledge and technology into the cluster system; group through the formation of cross-regional and cross-border alliance organizations; carry out cross-border regional exchanges and cooperation; organize and coordinate external network contacts; and promote the interaction and exchange of innovative resources. On the one hand, we need to guide large infrastructure equipment companies to promote exchanges and cooperation through direct investment; policy support, and innovation platforms; stimulate innovative production; and narrow the knowledge gap with the outside to improve technological capabilities. On the other hand, through industrial policies, fiscal policies, land policies, and copyright policies, large-scale equipment manufacturers are motivated to disseminate knowledge to other entities in the park.

Industrial systems have the structural characteristics of multidimensional heterogeneous nesting, and each scale contains the heterogeneity of elements and the heterogeneity of structure, environment, and system functions. This paper uses the framework of system theory to construct a variety of heterogeneity and study its mechanism in the evolution of industrial systems. The improvement of quality and efficiency is the ultimate goal of the development and upgrading of the equipment manufacturing industry. The quality of industrial inputs is critical to the improvement of industrial production efficiency. This paper uses industrial input quality indicators to study the impact of quality differences on the evolution of green total factor productivity in equipment manufacturing. Using quality indicators for input and output to achieve logical consistency can solve the relationship between quality and quality and make quality analysis more accurate.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71771083) and Research Projects of Hunan Provincial Department of Education of China (20A298).

References

[1] H.-L. Guan, "An evolution model for regional collaborative innovation under the perspective of complex network," *Journal of Intelligent & Fuzzy Systems*, vol. 31, no. 3, pp. 1319–1328, 2016.

[2] F. Piccinno, R. Hischier, A. Saba, D. Mitrano, S. Seeger, and C. Som, "Multi-perspective application selection: a method to identify sustainable applications for new materials using the example of cellulose nanofiber reinforced composites," *Journal of Cleaner Production*, vol. 112, no. 1, pp. 1199–1210, 2016.

[3] X. Li, "The innovative management mechanism for the ecological environment of photovoltaic new energy industrial clusters," *Light and Engineering*, vol. 25, no. 3, pp. 37–43, 2017.

[4] D. V. Hertem, W. Leterme, G. Chaffey et al., "Substations for future HVdc grids: equipment and configurations for connection of HVdc network elements," *IEEE Power and Energy Magazine*, vol. 17, no. 4, pp. 56–66, 2019.

[5] A. E. Karlik and V. V. Platonov, "Cross-industry spatially localized innovation networks," *Economy of Region*, vol. 1, no. 4, pp. 1218–1232, 2016.

[6] Q. Zhong, "Evolution model and simulation of industrial cluster innovation network based on complex network," *C e Ca*, vol. 42, no. 2, pp. 740–745, 2017.

[7] J. Jin-Bo and P. Xu-Dong, "Leakage and stiffness characteristics of bionic cluster spiral groove dry gas seal," *Chinese Journal of Mechanical Engineering*, vol. 31, no. 2, pp. 148–158, 2018.

[8] F. Chang, G. Zhou, X. Xiao, C. Tian, and C. Zhang, "A function availability-based integrated product-service network model for high-end manufacturing equipment," *Computers & Industrial Engineering*, vol. 126, pp. 302–316, 2018.

[9] R. C. Yam, "Does country-level R&D efficiency benefit from the collaboration network structure?" *Research Policy: A Journal Devoted to Research Policy, Research Management and Planning*, vol. 45, no. 4, pp. 770–784, 2016.

[10] Y. Lu, G. Shailer, and M. Wilson, "Corporate political donations: influences from directors' networks," *Journal of Business Ethics*, vol. 135, no. 3, pp. 461–481, 2016.

[11] Z. Ke, "Research on hemingway inferiority complex from the perspective of Adler's compensation mechanism," *Agro Food Industry Hi Tech*, vol. 28, no. 1, pp. 764–767, 2017.

[12] H. E. Fan, J. M. Zhang, and T. H. Gao, "Evolution mechanism of the topological structure during solid-liquid phase transition of InGaAs crystal," *Chinese Ence Bulletin*, vol. 62, no. 7, pp. 693–699, 2017.

[13] J. L. Tangen and C. S. Cashwell, "Touchstones of connection: a concept mapping study of counselor factors that contribute to relational depth," *The Journal of Humanistic Counseling*, vol. 55, no. 1, pp. 20–36, 2016.

[14] S. V. Dronov and E. A. Evdokimov, "Post-hoc cluster analysis of connection between the forming characteristics," *Model Assisted Statistics and Applications*, vol. 13, no. 2, pp. 183–195, 2018.

[15] D. Dilawaer and Hiroki, "Simple algorithms for selecting an energy-efficient server in a cluster of servers," *International Journal of Communication Networks and Distributed Systems: IJCNDS*, vol. 21, no. 1, pp. 1–25, 2018.

[16] Y. Hong, P. Can, Y. Xiaona, and L. Ruixue, "Does change of industrial structure affect energy consumption structure: a study based on the perspective of energy grade calculation," *Energy Exploration & Exploitation*, vol. 37, no. 1, pp. 579–592, 2019.

[17] S. Kergroach, "National innovation policies for technology upgrading through GVCs: a cross-country comparison," *Technological Forecasting and Social Change*, vol. 145, pp. 258–272, 2019.

[18] S. Fraser-Bell, R. Symes, and A. Vaze, "Hypertensive eye disease: a review," *Clinical & Experimental Ophthalmology*, vol. 45, no. 1, pp. 45–53, 2017.

[19] W. Gao and Z. Zhu, "The technological progress route alternative of carbon productivity promotion in China's industrial sector," *Natural Hazards*, vol. 82, no. 3, pp. 1803–1815, 2016.

[20] S. Tiley Jaimie, "A century of progress at the materials and manufacturing directorate throughout its 100-year history, the Materials and Manufacturing Directorate (ML) has conducted groundbreaking critical research programs responsible for a multitude of technological advances," *Advanced Materials & Processes*, vol. 175, no. 2, pp. 24–27, 2017.

[21] A. K. Yetisen, A. F. Coskun, G. England et al., "Art on the nanoscale and beyond," *Advanced Materials*, vol. 28, no. 9, pp. 1724–1742, 2016.

[22] J. Li, K. F. See, and J. Chi, "Water resources and water pollution emissions in China's industrial sector: a green-biased technological progress analysis," *Journal of Cleaner Production*, vol. 229, pp. 1412–1426, 2019.

[23] T. Mertelmeier and H. M. Hofmann, "Consistent cluster model description of the electromagnetic properties of lithium and beryllium nuclei," *Nuclear Physics*, vol. 459, no. 2, pp. 387–416, 2016.

[24] H. Chen, P. He, C. X. Zhang et al., "Efficiency of technological innovation in China's high tech industry based on DEA method," *Journal of Interdisciplinary Mathematics*, vol. 20, no. 6-7, pp. 1493–1496, 2017.

Hindawi
Complexity
Volume 2021, Article ID 6659302, 14 pages
https://doi.org/10.1155/2021/6659302

Research Article

Use of BP Neural Networks to Determine China's Regional CO_2 Emission Quota

Yawei Qi, Wenxiang Peng, Ran Yan, and Guangping Rao

School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330032, China

Correspondence should be addressed to Yawei Qi; qiyawei@jxufe.edu.cn

Received 26 November 2020; Revised 4 December 2020; Accepted 14 December 2020; Published 6 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

China declared a long-term commitment at the United Nations General Assembly (UNGA) in 2020 to reduce CO_2 emissions. This announcement has been described by Reuters as "the most important climate change commitment in years." The allocation of China's provincial CO_2 emission quotas (hereafter referred to as quotas) is crucial for building a unified national carbon market, which is an important policy tool necessary to achieve carbon emissions reduction. In the present research, we used historical quota data of China's carbon emission trading policy pilot areas from 2014 to 2017 to identify alternative features of corporate CO_2 emissions and build a backpropagation neural network model (BP) to train the benchmark model. Later, we used the model to calculate the quotas for other regions, provided they implement the carbon emission trading policy. Finally, we added up the quotas to obtain the total national quota. Additionally, considering the perspective of carbon emission terminal, a new characteristic system of quota allocation was proposed in order to retrain BP including the following three aspects: enterprise production, household consumption, and regional environment. The results of the benchmark model and the new models were compared. This feature system not only builds a reasonable quota-related indicator framework but also perfectly matches China's existing "bottom-up" total control quota approach. Compared with the previous literature, the present report proposes a quota allocation feature system closer to China's policy and trains BP to obtain reasonable feature weights. The model is very important for the establishment of a unified national carbon emission trading market and the determination of regional quotas in China.

1. Introduction

Because of the 2020 COVID-19 epidemic, human beings have become more aware of their relationship with nature and of the importance of sustaining a harmonious coexistence of man and nature. In a time of significant crises, including the COVID-19 epidemic and climate change, the international community agrees that only through the development and implementation of green and low-carbon technologies, society can achieve high-quality economic recovery [1–3]. On Sept. 22, during the General Debate of the 75th Session of the UNGA, Chinese President Xi Jinping declared that China aims to reach CO_2 emissions peak before 2030 and achieve carbon neutrality before 2060. According to this, China plans to restore its economy by promoting low-carbon technologies and lifestyle. Reducing CO_2 emissions has become an important goal of China's 14th Five-Year Plan.

During the "13th Five-Year Plan" period, the Chinese government learned from the successful experience of the European Union, Japan, and other economies in reducing CO_2 emissions and began to explore the use of market-based methods: CO_2 emission trading systems [4, 5]. From 2014 to 2019, the central government implemented a pilot CO_2 emission trading policy in 7 provinces and cities, and each regional government formulated relevant trading standards and rules. The government in the pilot regions implemented the "allocation + trading" quota management rules for emission-control enterprises, that is, the emission-control enterprises received free quotas issued by the government at the beginning of the performance period. These quotas are determined after an enterprise's self-inspection and CO_2 emissions report is issued and a third-party verification is performed. If an enterprise's CO_2 emission is exceeded/remained during the period, it can be bought/sold in the

carbon trading market. After six years, China's pilot carbon market has increased and has become the world's second largest carbon market in terms of quota trading volume. Preliminary statistics have shown that a total of 2,837 emission-control agencies, 1082 nonemission-control agencies, and 11,169 individuals have participated in the pilot market. The cumulative transaction volume of the 7 pilot market quotas has reached 406 million tons, and the cumulative transaction volume is about 9.28 billion yuan. By the end of 2019, China's carbon intensity was reduced by about 48.1% as compared with data from 2005. In addition, nonfossil energy accounted for 15.3% of the primary energy consumption, which means that China has achieved its 2020 emission reduction target ahead of schedule. Thus, China is currently applying market mechanisms to control and reduce greenhouse gas emissions and to promote the green and low-carbon transformation of economic development. Moreover, the implementation of the CO_2 emissions trading market represents not only an important institutional innovation for China but also an important policy tool to implement international agreements for emissions reduction. Given the significant emissions reduction results due to the implementation of the CO_2 emission trading market policies in the pilot regions, the Chinese government has initiated the creation of a unified national CO_2 emission trading market to help all regions in the country to reduce carbon emissions. It is expected that with these new rules, the CO_2 emission peak target by 2030 will be achieved as soon as possible. Therefore, the design and implementation of a unified national CO_2 emission market is an issue that needs to be studied urgently. The foundation for a proper design and implementation of a CO_2 emission trading market program to achieve the intended emission reductions resides in the correct determination of the national quota and quotas for each region (province and city).

After finding alternative features of corporate carbon emissions, we used quota data on China's carbon emission trading policy pilot areas from 2014 to 2017 and the BP model to calculate the quotas of other regions in the sample interval, provided that such regions implemented the carbon emissions trading policy. We obtained the national total quota. With respect to carbon emission terminal, we divided the quota allocation system into three aspects: enterprise production, household consumption, and regional environment and then retrained the BP to obtain new results which were compared to those of the benchmark model. We found out that (1) from 2014 to 2016, China's total quota displayed a yearly increase and a sudden decrease in 2017. During the initial stage of the national CO_2 emission trading market program, the national quota is expected to maintain a relatively high level. Later, during the following 3 to 4 years, through adjustment and adaptation, quotas in each region are expected to show a downward trend and increase in the change rate. This forces enterprises to either participate in the CO_2 emission trading market or improve their technology to reduce emissions. (2) Considering the feature system built by adding household consumption and regional environment, the training model displays a smaller loss rate, and the test results (other regional quotas) describe the actual situation in a more accurate way. Thus, when building a unified national CO_2 emission trading market and determining quotas for various regions in addition to enterprise production, it is more reasonable to consider a feature system that takes into account household consumption and regional environment. At the same time, this feature system can be used in combination with China's "bottom-up" total control and postadjustment method. It not only allows regional quota decision makers to predict CO_2 emissions in advance through the existing data in the feature system and the trained model before obtaining the final real summary of the CO_2 emissions but also allows enterprises considering CO_2 price when making investment decisions and trying to make profits in the market.

2. Literature Review

Previous research has mainly studied the following two aspects: (a) initial distribution of quotas for CO_2 emissions according to different principles and methods and (b) the allocation efficiency according to the initial allocation of carbon emission quotas.

The publications in the first area reported the study of distribution subjects and distribution methods. CO_2 emission quota allocation can be divided according to the two following perspectives: region and industry. Region includes the initial quota allocation among countries and the initial quota allocation among different provinces within the same country. In this context, the earliest research studied CO_2 emission quota allocation among countries. In 1992, the United Nations Conference on Environment and Development established "common but differentiated responsibilities" as the principle of international environmental cooperation. Later, in 1998, Rose et al. introduced the principle of CO_2 emission rights allocation, which should include equity and efficiency. However, what kind of allocation method involves "equity" and how to balance the importance of "equity" and "efficiency" are controversial topics [6]. Some scholars in developed countries believe that allocating CO_2 emission rights based on population size is in line with the principle of equity [7]. However, other experts in developing countries have proposed that "accumulated emissions per capita" is more related to equity. According to this concept, the allowed CO_2 emissions per capita in developing countries during the development stage should be higher than those in developed countries [8, 9]. Later, more researchers used different features and methods to quantitatively analyze the initial quota [10–12]. Based on population, GDP, and CO_2 emission data for 132 countries, Wang et al. proposed the Gini coefficient optimization model that optimizes the historical CO_2 emissions quota for various countries and is able to project future quotas [13].

Chinese scholars have focused on CO_2 quota allocation among provinces and industries in China [14–19]. Song et al. considered the comprehensive distribution principle of three indicators: hereditary system, egalitarianism, and payment, to create the provincial environmental fixed cost allocation optimization model (FCAM), which will be able to determine the allocation of provincial CO_2 emission

rights for 2020, with a more balanced equity and efficiency [20]. From the perspective of equity and efficiency, Yu and Wu used a master-slave hierarchical interactive iterative algorithm based on satisfaction to build a two-level planning model (the upper-level planning model based on equity and the lower-level planning model based on efficiency) to optimize the allocation of CO_2 emission rights between provinces [21]. Qian et al. used Chinese enterprise carbon patent data, and from a consumption and production perspective created the stochastic frontier model to measure regional CO_2 emission efficiency. Later, they used the estimated efficiency value to numerically simulate the regional allocation of CO_2 emission rights [22]. Wu et al. proposed a coupling model of China's multiregional CGE and CO_2 trade (CGE-3MS). The model showed the decision-making and optimization process of trading CO_2 units, and they analyzed the impact of the carbon market on the economy and emission-control industries in China under different initial quotas [23].

Compared with the first piece of literature, the second one focused on whether the existing allocation of quotas between regions and industries is reasonable and effective. In general, allocation efficiency is used to measure whether the allocation is reasonable. Some Chinese scholars used the original DEA and the zero-sum DEA models to measure quota allocation for Chinese provinces. They found out that the results of the zero-sum DEA model were better than those obtained with the original DEA model [24]. In addition, other scholars revised the original DEA model and proposed a new allocation method that included the evaluation of DEA efficiency and historical CO_2 emissions. Later, they used China's emission commitment as the decision-making guidance and selected the maximization of the average efficiency as the final goal [25, 26]. Besides, after proving that the original DEA model was inefficient to determine CO_2 emissions quotas, some scholars used other models to redistribute the quotas of various industries [27–29]. Huang and Zhang used the SBM and RE/CE models to get a more comprehensive efficiency that reveals Chinese energy use and the CO2 emission situation. From the empirical study of 30 regions in China, they found that the southern region of China has the most efficient score, while northeastern China has poor performance. Price factor has a significant influence on energy use and CO_2 emission efficient score of some provinces [30].

In summary, we have identified two points that should be taken into account when creating regional quota allocation systems and building the corresponding models. First, feature selection should be performed with caution. Factors related to the regional CO_2 emission accounting as well as those related to CO_2 emissions of the accounting entities (enterprises participating in the CO_2 emission trading market) should be selected. Second, when using the feature system data in the allocation model, it is necessary to carefully determine the weight of each feature to ensure the scientific, rigorous, and accurate quota allocation.

Therefore, in the present research, we first selected features that are closely related to the accounting entities (emission-control enterprises), such as the number of industrial enterprises exceeding designated size in the region, the energy consumption per unit of industrial added value, and the proportion of coal-fired energy included in the total energy consumption. We also considered features that are related to regional accounting CO_2 emissions, such as per capita energy consumption, per capita carbon emissions, forest coverage, and green areas. Then, we built a feature system that can be used in conjunction with China's "bottom-up" total control and postadjustment method to provide a predictive model. In addition, this model allows enterprises considering CO_2 prices when making investment decisions, thereby stimulating corporate green innovation. In addition, for the first time, we considered the use of the BP in order to determine the weight of each feature based on historical quota data. This is different from the previous literature, which used subjective weight determination methods. In addition, our methods are more in line with China's actual CO_2 emission situation and more accurate during calculations. Finally, the feature system and the corresponding BP neural network model proposed in the present research can be used to calculate the quotas of other regions during the same period and also predict future quotas for the same region.

3. Empirical Analysis

3.1. Data and Feature Statistics

3.1.1. Data. On October 29, 2011, China´s Development and Reform Commission indicated that China should start implementing the pilot carbon emission trading policy. Specifically, Beijing, Tianjin, Shanghai, Chongqing, Guangdong, Hubei, and Shenzhen were labeled as the pilot regions. This allowed each regional government to determine the transaction start time, corresponding transaction subject enterprises, and quotas according to local conditions. Although the central government promulgated the pilot policy in 2011, each pilot region actually started the carbon emission trading system between 2013 and 2014. Therefore, considering the complete CO_2 quota and the availability of additional data, we selected 31 regions (5 cities and 26 provinces) as the research sample for the period 2014–2017. With respect to sample pretreatment, the following aspects were considered: (1) given that Shenzhen belongs to Guangdong Province and that both of them are pilot regions, Shenzhen was analyzed separately; (2) regions with serious missing values in historical feature data were eliminated (Tibet); (3) taking into account that the different dimensions of the data may affect the prediction, all features were normalized using

$$b_{ijt} = \frac{a_{ijt} - \min a_{ijt}}{\max a_{ijt} - \min a_{ijt}}, \tag{1}$$

where a_{ijt} indicates the jth original feature of the i region in the t year, $\min a_{ijt}$ and $\max a_{ijt}$ represent the minimum and maximum values of the original feature j in 31 regions in the t year, respectively, and b_{jit} indicates the j feature of i region in the t year after normalization.

Regional features (except that for Shenzhen) were obtained from the National Bureau of Statistics and China Statistical Yearbook. Features of Shenzhen were obtained from the Shenzhen Statistical Yearbook, and the historical quotas data came from the 2014–2018 Beijing Carbon Market Annual Report.

China and its regions have not released official data on CO_2 emissions. Thus, we used the CO_2 emission calculation method given in the 2006 IPPC National Greenhouse Inventory Guidelines and reported by Qi [31]:

$$\begin{aligned} CO_2 &= \sum_{i=1}^{n} EF_i \times E_i \times 10^{-6} \\ &= \sum_{i=1}^{n} CEF_i \times NCV_i \times OR_i \times \left(\frac{44}{12}\right) \times E_i \times 10^{-6}, \end{aligned} \tag{2}$$

where CO_2 indicates carbon dioxide emissions (t) and EF_i represents the carbon dioxide emissions factor for a specific fuel ($kg - CO_2/kg, m^3$). We considered 11 types of energy consumption including raw coal, coal, coke, crude oil, fuel oil, gasoline, kerosene, diesel, liquefied petroleum gas, refinery dry gas, and natural gas ($n = 11$); E_i is the fuel consumption for fuel i (kg, m^3); CEF_i represents the carbon content of the fuel i (tC/TJ); NCV_i indicates the average low calorific value of fuel i ($kJ/kg, m^3$); OR_i denotes the carbon oxidation rate of fuel i (%). The carbon dioxide emission coefficients of various energy sources are shown in Table 1.

Table 1: The CO_2 emission coefficients of various energy sources.

Energy	NCV	CEF	OR	EF
Raw coal	20 908	26.37	0.94	1.900 3
Coal	20908	25.8	0.92	1.8300
Coke	28435	29.5	0.93	2.860 4
Crude oil	41816	20.1	0.98	3.020 2
Fuel oil	41816	21.1	0.98	3.170 5
Gasoline	43070	18.9	0.98	2.925 1
Kerosene	43070	19.5	0.98	3.017 9
Diesel	42652	20.2	0.98	3.095 9
Liquefied petroleum gas	50179	17.2	0.98	3.101 3
Refinery dry gas	46055	18.2	0.98	3.011 9
Natural gas	38931	15.3	0.99	2.162 2

Notes: (1) NCV comes from China "General Principles of Comprehensive Energy Consumption Calculation" (GB/T 2589–2008); (2) CEF and OR come from China's "Provincial Greenhouse Gas Inventory Compilation Guide" (NDRC Office [2011] No. 1041).

3.1.2. Feature Statistics. Considering the calculation order, there are two carbon allocation methods:

(1) The "up-bottom" allocation method is applied from a macro (provincial) perspective, according to the general principle of efficiency and equity. This methodology considers population, provincial economic level (GDP), degree of industrialization (industrial structure), historical factors (accumulated carbon emissions per capita), and natural endowments to determine quotas for the different regions and for enterprises. In addition, the method allows determining the amount of quotas in advance, in such a way that participating entities are able to consider CO_2 prices in decision-making processes. However, due to the lack of data regarding actual emissions of the participating entities, when the quotas correspond to the actual situation and, in consequence, entities are motivated to participate in market transactions emissions reduction, these processes display great uncertainty.

(2) In the case of the "bottom-up" allocation method, the level-by-level summary determines the quotas for each region through the calculation of emissions at the microlevel (emission terminal). Considering carbon dioxide production terminal in human society (mainly consumers and enterprises), the enterprise component considers the number of equipment that emits CO_2, the scale of the enterprise, the scientific research level of enterprise emissions reduction (patents and R&D investment), and so on. The consumer component involves population size, per capita carbon emissions, and so on. Taking into account the emission data of the participating entities, this method is suitable for adjusting quotas afterwards (without specifying quotas in advance). However, it may eventually result in oversupply of quotas due to the excessively high emission limits. This situation may discourage participants to consider CO_2 prices when they are making investment decisions. In consequence, participants are not motivated to reduce emissions.

At the same time, given the problem of excessive CO_2 emissions caused by humans, people are also planting trees (forest carbon sinks) and using photosynthesis to reduce emissions. Therefore, resident life should also be considered when calculating quotas.

In 2020, the central government will formally begin to build a national carbon emission trading market. First, it will issue a national carbon emission trading quota setting and allocation implementation plan for the power industry (2019–2020). The plan involves a "bottom-up" quota determination method, that is, relevant provincial departments will be required to determine the list of key emitters and their actual output. Later, they will identify key emitters' quotas based on the benchmark method (the free quotas in each region for 2019–2020 were preallocated according to the 70% of the power (heat) supply of each key emitter in 2018). Then, after the quotas of all the key emitters in each region are verified, they will be added up to form the total quota of the region, and the regional quotas will be further added to obtain the total quota of the country.

According to China's "bottom-up" quota allocation method, it is most accurate to use the CO_2 emission equipment data of emission-control enterprises in the regions. However, since China's national carbon emission trading market is still in its infancy, relevant data (corporate power supply (heat) units, actual output) of emission-control enterprises in regions have not been released.

Also, "bottom-up" quota allocation methodologies present several limitations, which may result in oversupply of quotas, insufficient demand, low CO_2 price, and inactive market due to excessively high emission limits. Because of this, participants usually do not consider CO_2 prices when they are in the process of investment decision-making and cannot effectively motivate participants to reduce emissions.

Therefore, based on the operating experience of The European Union Emission Trading Scheme (EU ETS) and Regional Greenhouse Gas Initiative (RGGI), we proposed a more reasonable "top-down" allocation method in combination with China's existing "bottom-up" quota allocation method [32]. This method was divided into two steps. The first step involved the development of a feature system that is more suitable for China's provincial carbon quota allocation. This was performed by selecting features that are related to participating entities. For example, (a) the current emission-control enterprises in China are mainly industrial enterprises; thus, the region that contains more industrial enterprises will have more quotas; (b) energy consumption, including coal-fired energy that is used by emission-control enterprises in production activities ranks first in regional energy consumption; thus, energy structure and energy consumption per unit GDP are both factors affecting regional quotas. The second step involves training suitable BP to calculate nonpilot regional quotas based on historical pilot regional quotas and feature system data.

Most ETS are based on total CO_2 emission control. Determination of total CO_2 emissions should not only consider the overall emission target, but also the regional differences (level of economic development, technological differences, and forest carbon sink).

Based on previous reports and considering different factors included in the "top-down" quota control approach adopted by the EU ETS after phase II [33–35], we divided the features into three categories: (a) enterprise production, (b) household consumption, and (c) regional environment.

(a) Household consumption: in 2019, 30% of China's CO_2 emissions corresponded to consumer and man-made emissions. Household consumption, as the main body of society, is one of the main causes of CO_2 emissions. Regions with large populations display frequent economic activities. Therefore, resident consumption factors depend on per capita GDP, energy consumption, carbon emissions, cumulative carbon emissions, and disposable income.

(b) Enterprise production: in 2019, 70% of China's CO_2 emissions were the result of industrial production or generative emissions. Among them, the carbon dioxide emissions related to the power industry accounted for more than 40%. In addition, those from the steel industry, which are part of the manufacturing industry, accounted for about 15%. Industrial enterprises exceeding permitted size are not only the main body of the industry but also the main entities participating in the carbon emissions trading market. Therefore, factors related to CO_2 emissions from enterprises include the number of industrial enterprises exceeding permitted size, full-time equivalent of R&D personnel in industrial enterprises exceeding designated size, industrial structure, and energy structure, energy consumption per GDP unit, energy consumption per unit of industrial added value, and electricity consumption per GDP unit.

(c) Regional environment: in addition to factors related to resident and enterprises, regional environment also affects CO_2 emissions and decomposition. We considered three aspects, regional economic level, technological level, and green resources, mainly including total freight volume, total passenger volume, degree of openness to the outside world, total gas supply, total supply of liquefied petroleum gas, urbanization rate, technological level, green area, and forest carbon sink. Specific explanations of features are given in Table 2.

3.2. Empirical Model

3.2.1. Specifications of the Backpropagation Neural Network Model (BP). Compared with the subjective weight assignment methods (AHP, Expert Evaluation Method, TOPSIS, etc.) used in the formation of the "top-down" model and reported in previous literature, BP involves a multilayer feedforward network and error direction propagation-learning algorithm. Because of its unique adaptability, learning ability, and strong generalization ability, it is widely used in the fields of automatic identification, predictive estimation, engineering, biology, and medicine, among others. For the purpose of the present research, BP can more objectively and accurately quantify the impact of features on quotas and dynamically reflect the nonlinear impact of features on quotas at different stages [36]. Therefore, after training the BP based on pilot regional quotas and feature system data, we calculated the nonpilot regional quotas.

BP is composed of an input layer, a hidden layer, and an output layer. These three basic elements are fully connected during the whole network training. For example, for a neural network model with only one hidden layer, the process of BP neural network is mainly divided into two stages.

The first stage is the forward propagation of the signal, from the input layer to the hidden layer and finally to the output layer. Assuming that the number of samples is A, input layer has m nodes, output layer has n nodes, and hidden layer has p nodes; x_{ai} is the input/output of the input layer, $a = 1, 2, \ldots, A$, $i = 1, 2, \ldots, m$; B_{aj}, b_{aj} is the input/output of the hidden layer, respectively, $j = 1, 2, \ldots, p$; Y_{ak}, and y_{ak} are the input/output of the output layer, respectively, $k = 1, 2, \ldots, n$; y^*_{ak} is the expected label of the output layer (historical quotas); w_{ij} and w_{jk} are the weights from input to hidden and hidden to output, respectively; , θ_j and γ_k are the biases from input to hidden and hidden to output, respectively.

Table 2: Specific explanation of features.

Feature	Explanation
GDP per capita	GDP/population
Energy consumption per capita	Energy consumption/population
Degree of opening to the outside world	Total import and export/GDP
Carbon emissions per capita	Total CO_2 emissions/population
Cumulative carbon emissions per capita	Cumulative CO_2 emissions/Cumulative population
Total passenger volume	The number of passengers actually carried by the means of transport in a certain period of time
Disposable income per capita	Disposable income/population
Industrial structure	Added value of tertiary industry/GDP
Energy consumption per unit of GDP	Energy consumption/GDP
Energy consumption per unit of industrial added value	Energy consumption/industrial added value
Electricity consumption per unit of GDP	Electricity consumption/GDP
Energy structure	Coal-fired energy/energy consumption
Total freight volume	In a certain period of time, the actual weight of the cargo carried by the means of transport
The number of industrial enterprises above designated size	Including independent auditor industrial enterprises and affiliated industrial production units
Full-time equivalent of R&D personnel in industrial enterprises above designated size	The number of full-time staff plus part-time staff is converted to the total number of full-time staff based on workload
Technological level	Number of patents granted/10,000 people
Forest carbon sink	Forest area/total land area
Green area	Including park area, production green area, protective green area, subsidiary green area, and other green area
Total gas supply	Refers to the amount of natural gas supplied by gas units during the reporting period
Total supply of liquefied petroleum gas	Refers to the amount of liquefied petroleum gas supplied by gas units during the reporting period
Urbanization rate	Urban population/population

Input layer to hidden layer: determine input function, $B_{aj} = x_{ai}w_{ij} + \theta_j$, and then transform B_{aj} into b_{aj} through activation function $b_{aj} = f(B_{aj})$.

Hidden layer to output layer: determine input function, $Y_{ak} = b_{aj}w_{jk} + \gamma_k$, and then transform Y_{ak} into y_{ak} through activation function, $y_{ak} = f(Y_{ak})$, where y_{ak} is the final result.

Later, the loss function is determined and loss is calculated according to y_{ak} and y^*_{ak}. When the loss is either smaller than the set range or reaches the upper limit of the number of iterations, the model ends the training; otherwise, it enters a second stage.

The second stage is the backpropagation of the loss. The loss information is returned along the original propagation path through the learning signal. Starting from the last layer, the weight and bias are corrected layer by layer, and finally the loss is within the set range.

Weight and bias update formula:

$$\begin{aligned} W^{l*} &= W^l + \Delta W^l, \\ \theta^{l*} &= \theta^l + \Delta\theta^l, \\ \Delta W^l &= -\eta \frac{\partial \text{Loss}}{\partial W^l}, \\ \Delta\theta^l &= -\eta \frac{\partial \text{Loss}}{\partial \theta^l}, \end{aligned} \tag{3}$$

where W^{l*} and θ^{l*} indicate the updated weight and bias of the l layer, W^l and θ^l indicate the original weight and bias of the l layer, ΔW^l and $\Delta\theta^l$ represent the correction part of weight and bias of the l layer, and η is a fixed value that indicates the learning rate.

The fundamental part of backpropagation is to minimize the loss through the update of weights and biases, using the gradient descent method (actually using the chain partial derivative). The specific derivation process is given below.

The following assumptions are considered:

The activation function from the input layer to the hidden layer is $f(x) = (1/(1 + e^{-x}))$ (sigmoid).

The activation function from the input layer to the hidden layer is $f(x) = x$.

Loss function is $\text{Loss} = (1/2) \times \sum_{k=1}^{n} (y^*_{ak} - y_{ak})^2 = (1/2) \times \sum_{k=1}^{n} e^2_{ak}$, $e_{ak} = y^*_{ak} - y_{ak}$.

The weights update process of the hidden-output layer:

$$\begin{aligned} \frac{\partial Loss}{\partial w_{jk}} &= \frac{\partial Loss}{\partial e_{ak}} \times \frac{\partial e_{ak}}{\partial y_{ak}} \times \frac{\partial y_{ak}}{\partial Y_{ak}} \times \frac{\partial Y_{ak}}{\partial w_{jk}} \\ &= e_{ak} \times (-1) \times 1 \times b_{aj} = -b_{aj}e_{ak}, \\ w^*_{jk} &= w_{jk} - \eta \frac{\partial Loss}{\partial w_{jk}} = w_{jk} + \eta b_{aj}e_{ak}. \end{aligned} \tag{4}$$

The weight update process of the input-hidden layer:

$$\frac{\partial Loss}{\partial w_{ij}} = \frac{\partial Loss}{\partial e_{ak}} \times \frac{\partial e_{ak}}{\partial y_{ak}} \times \frac{\partial y_{ak}}{\partial Y_{ak}} \times \frac{\partial Y_{ak}}{\partial b_{aj}} \times \frac{\partial b_{aj}}{\partial B_{ij}} \times \frac{\partial B_{ij}}{\partial w_{ij}}$$

$$= e_{ak} \times (-1) \times w_{jk} \times b_{aj}\left(1 - b_{aj}\right) \times x_{ai}$$

$$= -b_{aj}\left(1 - b_{aj}\right) w_{jk} x_{ai} e_{ak},$$

$$w_{ij}^{*} = w_{ij} - \eta \frac{\partial Loss}{\partial w_{ij}} = w_{ij} + \eta b_{aj}\left(1 - b_{aj}\right) e_{ak} w_{jk} x_{ai}. \tag{5}$$

In the same way, the updated value of the bias can be obtained:

Hidden-output layer: $\gamma_k^{*} = \gamma_k + \eta e_{ak}$

Input-hidden layer: $\theta_j^{*} = \theta_j + \eta b_{aj}(1 - b_{aj}) w_{jk} e_{ak}$

3.2.2. Key Parameters. According to the model principle previously mentioned, there are 5 key parameters that determine the learning effect of the BP: activation function, loss function, learning rate, the number of hidden layer nodes, and gradient descent algorithm.

(1) Activation function:

For the BP both, the hidden layer and the output layer need to use an activation function.

For the hidden layer, the activation function is generally a nonlinear function. The reason for this is that, if the activation function is a linear function, the output is a linear combination of the input, which is equivalent to the effect of the no hidden layer (the hidden layer is invalid). The introduction of a nonlinear function as the hidden layer activation function makes the network more powerful, increases its ability to learn complex data, and reflects the nonlinear relationship between input and output. Therefore, we introduced four nonlinear activation functions that are widely used (Tables 3 and 4).

Although ReLU has two problems, it is currently the most commonly used activation function for BP. In addition, it is the default activation function used by most feedforward neural networks.

For the output layer, the choice of its activation function depends on whether the problem is a regression problem or a classification problem. In the event, it is a classification problem, the sigmoid activation function represents a good choice; for regression problems, a linear activation function is more appropriate.

(2) Loss function:

With regards to the problem and the output layer activation function to match different loss functions,

(1) Cross-entropy function: it is suitable for binary classification problems, and the output layer activation function is sigmoid

(2) Log-likelihood cost: it is suitable for multiclassification problems, and the output layer activation function is softmax

(3) Mean square error (MSE): it is suitable for regression problems, and the output layer activation function is a linear function

(3) Learning rate:

The learning rate value is an important part of the BP, which represents the speed of information accumulation in the neural network over time, and its value is between [0, 1]. Under ideal circumstances, we would start with a large learning rate and gradually reduce the speed until the loss value no longer diverges (if the learning rate is set too low, the training progress will be very slow because only very few adjustments to the weight of the network are made. However, if the learning rate is set too high, it may bring undesirable consequences on the loss function (Figure 1)).

(4) The number of hidden layer nodes:

The number of hidden layer nodes has a great influence on the prediction accuracy of BP; if the number of nodes is too small, the network cannot perform a proper learning process, and it will need more times to train. In addition, the training accuracy is also affected. When the number of nodes is too large, the training time increases and the network will result in overfitting. However, there is no conventional formula for determining the number of nodes. Some empirical formulas are given below for reference: $l < n - 1$ $l < \sqrt{(m+n)} + a$, $l = \log_2 n$, $l \geq k \times n/(n+m)$, where n indicates the number of input layer nodes, l indicates the number of hidden layer nodes, m is the number of output layer nodes, a represents a constant between 0–10, and k corresponds to the number of samples.

In fact, the number of hidden layer nodes can be roughly calculated according to the reference formula. Later, trial and error is used to find the optimal number of nodes. Generally speaking, the BP error shows a trend where it first decreases and later increases with the increase of hidden layer nodes.

(5) Gradient descent algorithm:

We introduce six well-known gradient descent algorithms (Tables 5 and 6).

According to this analysis, there are no perfect key parameters that can suit all conditions. The appropriate selection of key parameters depends on the specific problem of study. In the present research, we studied a regression problem. Thus, we chose ReLU and $f(x) = x$ as the activation function and MSE as the loss function. Also, Adam may be appropriate as the gradient descent algorithm; however, the learning rate and the number of hidden layer nodes cannot be determined in advance. In summary, the final determination of all key parameters needs BP

Table 3: Activation functions and figures.

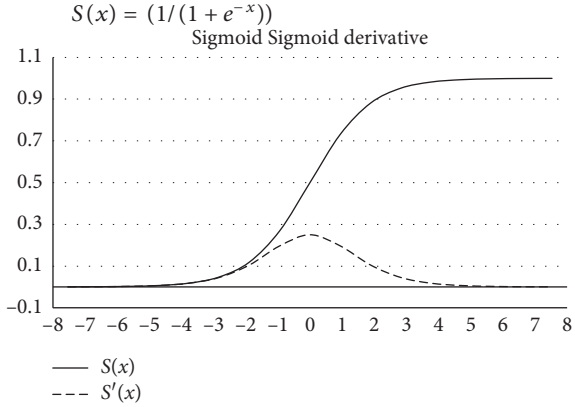
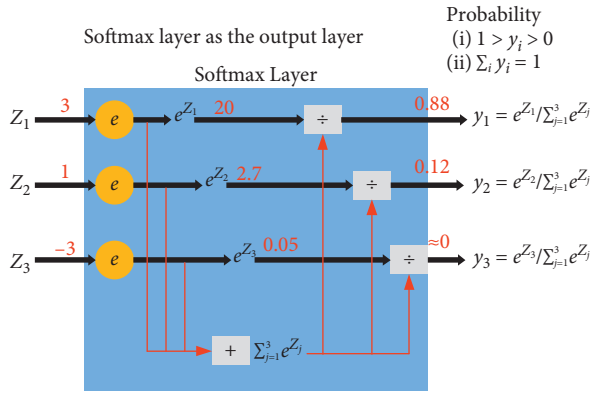
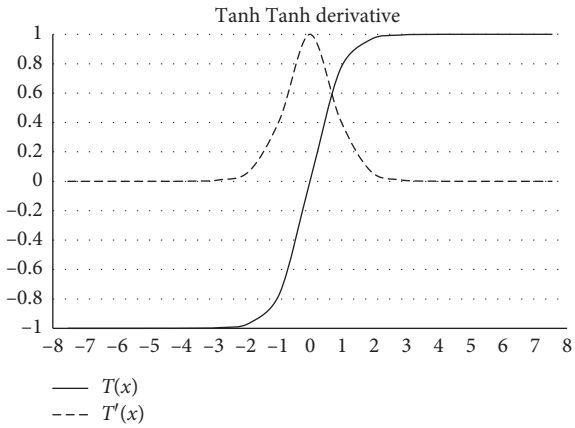
Name	Function	Derivative function
Sigmoid	$S(x) = (1/(1+e^{-x}))$ 	$S'(x) = S(x)(1 - S(x))$
Softmax	$\text{Softmax}(Z_i) = (e^{Z_i}/\sum_{c=1}^{C} e^{Z_c}) = p_i$ 	$(\partial \text{softmax}(Z_i)/\partial Z_j) = \begin{cases} p_i(1-p_i), & j = i, \\ -p_j \cdot p_i, & j \neq i. \end{cases}$
tanh	$\tanh(x) = ((e^x - e^{-x})/(e^x + e^{-x}))$ 	$T'(x) = 1 - (T(x))^2$

Table 3: Continued.

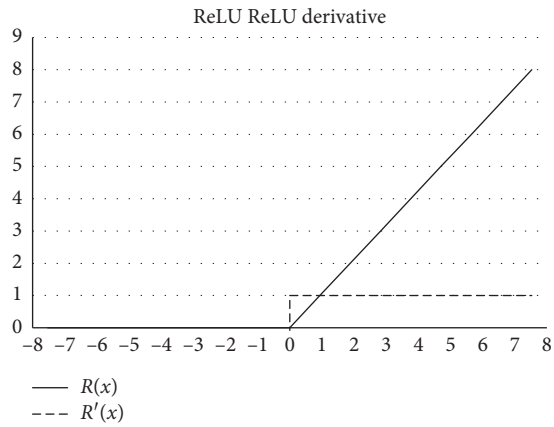
Name	Function	Derivative function
ReLU	$\mathrm{ReLU}(x) = \max(0, x)$ 	$\mathrm{ReLU}'(x) = \begin{cases} 0, & x \le 0, \\ 1, & x > 0. \end{cases}$

Table 4: Advantages and disadvantages of activation functions.

Name	Advantages	Disadvantages
Sigmoid	(1) It can smoothly map the real-number field to [0, 1] (2) Monotonically increasing, continuous derivable, and its derivative form is very simple (3) Suitable for handling binary classification problems	(1) Gradient vanishing, that is, in the process of backpropagation, the derivative will gradually become 0; thus, the parameters cannot be updated and the neural network cannot be optimized (2) Nonzero-centered: the output value of the function is always greater than 0, which will slow down the convergence speed of the model training (3) Exponentiation is relatively time-consuming
Softmax	(1) It maps the output value to (0, 1), and the sum of the mapped output value is 1 (2) It divides the entire hyperspace according to the number of classifications (3) Suitable for multiclassification problems	(1) The operation of Softmax involves the calculation of exponential function; in consequence, an "overflow problem" for computers occurs (2) Not suitable for face recognition tasks
tanh	(1) It can smoothly map the real-number field to [−1, 1] (2) Solve nonzero-centered problem (3) Suitable for handling binary classification problems	(1) Gradient vanishing, that is, in the process of backpropagation, the derivative will gradually become 0; thus, the parameters cannot be updated and the neural network cannot be optimized
ReLU	(1) Solve the gradient vanishing in the positive interval (2) The calculation is simple, no exponential calculation is required, and the activation value can be obtained with only one value (3) The convergence speed is much faster than sigmoid and tanh	(1) Nonzero-centered (2) Dead ReLU problem: it is "vulnerable" during training; when $x < 0$, the gradient is 0; the gradient of these nodes and subsequent nodes are always 0, and no longer responds to any data, causing the corresponding parameters to never be updated

training. For this reason, the determination of key parameters is provided in Section 4.

4. Empirical Results and Analysis

We focused on the allocation of regional carbon quotas, which is a regression problem, and our goal was to minimize the loss rate while ensuring that the test set results meet the realistic expected range in China. Therefore, based on experience and BP training, we chose MSE as the loss function and Adam [37] as the gradient descent algorithm in the backpropagation process, the final learning rate was 0.009, and number of iterations were 5000. Other parameters are shown in Table 7.

The total loss rate of the benchmark model was 0.02419. And the comparison between the results of the training set and the historical quotas are shown in Table 8 (benchmark model in Table 8). The test results are shown in Table 9 (benchmark model in Table 9). After adding up the historical quotas in pilot regions and the estimated quotas in the nonpilot regions, national quotas were obtained. The resulting national quotas are displayed in Table 10 (benchmark model in Table 10).

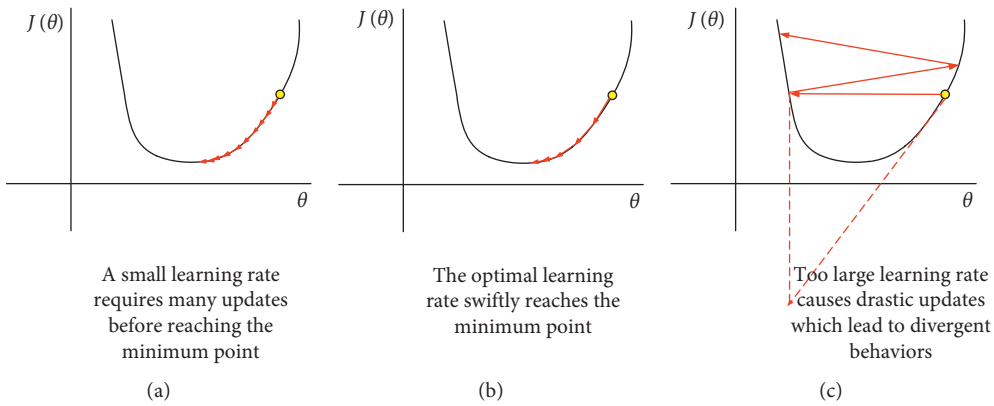


Figure 1: The loss function results of different learning rates.

Table 5: Gradient descent functions and parameter description.

Name	Gradient descent function	Parameter description
BGD	$\theta^* = \theta - \eta \cdot \widehat{g} = \theta - \eta \cdot \nabla_\theta J(\theta)$	$\widehat{g}$ denotes gradient estimate θ denotes weight
SGD	$\theta^* = \theta - \eta \cdot \widehat{g} = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$	$\widehat{g}$ denotes gradient estimate θ denotes weight
Momentum	$v_t = \gamma v_{t-1} + \eta \cdot \nabla_\theta J(\theta)$ $\theta^* = \theta - v_t$	vt denotes gradient estimate θ denotes weight
AdaGrad	$\widehat{g}_{t,i} = \nabla_\theta J(\theta_i)$ $\theta^*_{t+1,i} = \theta_{t,i} - (\eta / \sqrt{\widehat{g}_{t,i} \cdot \widehat{g}_{t,i} + \varepsilon}) \cdot \widehat{g}_{t,i}$	$\widehat{g}_{t,i}$ denotes gradient estimate $\theta_{t,i}$ denotes weight
RMSprop	$g_t = \nabla_\theta J(\theta_{t-1})$ $E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2$ $\theta^*_{t+1} = \theta_t - (\eta / \sqrt{E[g^2]_t + \varepsilon}) g_t$	$\gamma = 0.9 E[g^2]_t$ denotes Exponential weighted average g_t denotes gradient estimate θ_t denotes weight
Adam	$g_t = \nabla_\theta J(\theta_{t-1})$ $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ $\widehat{m}_t = (m_t / (1-\beta_1^t)),\ \widehat{v}_t = (v_t / (1-\beta_2^t))$ $\theta^*_{t+1} = \theta_t - (\eta / (\sqrt{\widehat{v}_t + \varepsilon})) \widehat{m}_t$	$\beta_1 = 0.9$ $\beta_2 = 0.999$ $\varepsilon = 10e-8$ g_t denotes gradient estimate θ_t denotes weight

Table 6: Advantages and disadvantages of gradient descent algorithms.

Name	Advantages	Disadvantages
BGD	The principle of gradient descent is simple	(1) Calculation is very slow (2) Difficult to handle a large dataset (3) Cannot add new data to update the model
SGD	(1) Compared with BGD, SGD training speed is faster (2) New data can be added to update the model	(1) Frequent updates may cause severe oscillations in the loss function
Momentum	(1) Consider the speed of the previous step and the new gradient (2) Can speed up the convergence and suppress the shock	
AdaGrad	(1) Compared with SGD, it adds a denominator (2) Handle the case where the number of gradient updates is small	(1) If gradient update is frequent, it may cause the subsequent gradient updates be slow or disappear
RMSprop	(1) Similar to momentum, it can reduce fluctuations (2) Overcome the problem of the sharp decrease or disappearance of gradient in AdaGrad (3) It performs better than SGD, momentum, and AdaGrad, based on the nonstationary objective function	
Adam	(1) Combine momentum and RMSProp (2) Integrate the contents of gradient descent, momentum, Adagrad, and RMSprop with certain improvement (3) Easy to use, insensitive to gradient scaling, can be used for large data, processing sparse data, easy to adjust hyperparameters, etc.	

Table 7: Parameters of the benchmark model.

Layer	Function	Activation
Import-hidden	(8, 6)	ReLU
Hidden-export	(6, 1)	$f(x) = x$

Table 8: Comparison between model results and historical quotas.

Year	Region	Historical quotas	Benchmark model		New model	
			Training results	Loss	Training results	Loss
2014	Beijing	0.5	0.5114	0.0114	0.502	0.002
	Tianjing	1.6	1.7397	0.1397	1.584	−0.016
	Shanghai	1.5	1.7795	0.2795	1.4785	−0.0215
	Hubei	3.24	3.0159	−0.2241	3.2485	0.0085
	Guangdong	3.88	3.6819	−0.1981	3.8754	−0.0046
	Shenzhen	0.33	0.3623	0.0323	0.3169	−0.0131
	Chongqing	1.3	1.3283	0.0283	1.2949	−0.0051
2015	Beijing	0.5	0.5441	0.0441	0.5144	0.0144
	Tianjing	1.6	1.7312	0.1312	1.6169	0.0169
	Shanghai	1.6	1.6728	0.0728	1.6214	0.0214
	Hubei	3.24	3.0162	−0.2238	3.1589	−0.0811
	Guangdong	4.08	3.8977	−0.1823	4.0595	−0.0205
	Shenzhen	0.33	0.3623	0.0323	0.3182	−0.0118
	Chongqing	1.25	1.419	0.169	1.3037	0.0537
2016	Beijing	0.5	0.4536	−0.0464	0.4882	−0.0118
	Tianjing	1.6	1.5162	−0.0838	1.5506	−0.0494
	Shanghai	1.5	1.4923	−0.0077	1.5099	0.0099
	Hubei	2.8	2.9384	0.1384	2.8715	0.0715
	Guangdong	4	4.0552	0.0552	4.0141	0.0141
	Shenzhen	0.3	0.3636	0.0636	0.3209	0.0209
	Chongqing	1.3	1.4176	0.1176	1.2564	−0.0436
2017	Beijing	0.5	0.3623	−0.1377	0.4946	−0.0054
	Tianjing	1.6	1.363	−0.237	1.6399	0.0399
	Shanghai	1.6	1.2582	−0.3418	1.5741	−0.0259
	Hubei	2.5	2.5891	0.0891	2.4883	−0.0117
	Guangdong	4.2	4.5011	0.3011	4.208	0.008
	Shenzhen	0.3	0.3623	0.0623	0.3043	0.0043
	Chongqing	1.3	1.2166	−0.0834	1.2883	−0.0117

Unit of quotas: 100 million tons.

Table 9: Test quotas of nonpilot regions.

Region	Benchmark model				New model			
	2014	2015	2016	2017	2014	2015	2016	2017
Hebei	4.4742	4.7213	4.6639	3.7958	4.2489	4.2512	3.9389	3.4056
Shanxi	8.7874	9.5514	9.9179	7.9321	6.7575	7.2639	7.4857	6.0955
Inner Mongolia	5.4541	5.3178	5.8367	6.0577	5.4135	5.2643	5.3196	5.8676
Liaoning	5.7035	5.2196	6.768	5.4917	5.9595	5.5083	5.8179	5.1618
Jilin	3.3295	3.1944	3.2286	2.8011	4.5732	4.651	4.5139	4.3528
Heilongjiang	4.5658	4.7666	5.3126	4.6711	5.1161	5.3088	5.5612	5.4386
Jiangsu	4.9033	4.9379	4.9276	4.5071	3.4851	3.6938	3.7096	3.6284
Zhejiang	4.3896	4.3216	4.0597	3.9379	4.2172	4.3214	4.0143	4.036
Anhui	3.6231	3.673	3.5841	3.1743	3.64	4.0833	3.8208	3.1997
Fujian	2.9149	2.9427	2.7986	2.5997	3.558	3.5671	3.3428	3.4086
Jiangxi	2.1524	2.3186	2.408	2.1446	3.5334	3.5535	3.4032	3.2111
Shandong	7.08	7.0777	7.2947	6.766	5.6471	5.6322	5.4154	5.0304
Henan	3.6061	3.7802	3.8308	3.2201	3.1582	3.2654	3.3632	2.937
Hunan	2.1693	2.3032	2.3678	2.1815	2.6993	2.8095	2.894	2.9341
Guangxi	2.0799	1.991	2.083	2.0383	3.2628	3.1519	3.0473	3.0321
Hainan	7.092	7.7761	7.7654	5.8029	6.1714	6.5658	6.6143	5.6134
Sichuan	3.0356	2.9589	2.8458	2.523	1.9639	2.1027	2.0326	1.9373

Table 9: Continued.

Region	Benchmark model				New model			
	2014	2015	2016	2017	2014	2015	2016	2017
Guizhou	3.3924	3.2882	3.5977	2.4839	3.4729	3.4067	3.4404	2.8913
Yunnan	1.1869	0.9986	1.1944	1.1038	2.698	2.6555	2.7258	2.5445
Shaanxi	4.4716	4.4987	4.5698	3.7151	2.7507	2.9126	2.689	2.6752
Gansu	3.9962	4.2983	4.357	3.2307	2.3042	2.6187	2.6927	2.0599
Qinghai	0.3623	0.3623	0.8758	0.3623	2.0744	1.9919	2.1466	2.0002
Ningxia	8.9951	9.0065	8.892	9.0385	4.9634	5.1076	5.0497	5.1049
Xinjiang	8.286	8.2076	9.1625	6.5578	3.6802	3.9403	4.236	3.8333

Unit of quotas: 100 million tons.

Table 10: National quota.

	National quota			
Benchmark model	2014	2015	2016	2017
	118.4012	120.1122	124.3424	108.137
New model	2014	2015	2016	2017
	107.6989	110.2274	109.2749	102.3993

Unit of quotas: 100 million tons.

The results of the benchmark model indicated that, during the initial stage of China's unified carbon emission market (3-4 years), the national quota will increase and, after an adaptation period, China's total quota and regional quotas will begin to decrease. This will stimulate enterprises to accelerate emission reduction and prove China's determination to achieve carbon peaks before 2030 and carbon neutrality by 2060.

5. Further Analysis: Build a Comprehensive Feature System

It is unreasonable to allocate regional quotas only considering the factor of corporate CO_2 emissions. In addition to corporate production factors, regional CO_2 emissions should also consider regional human activities and the role of forests in reducing those emissions. Thus, we believe that, in addition to the CO_2 emissions reported by enterprises, quotas in China's pilot regions should also take into account other features such as forest carbon sinks, population, and natural endowments. Based on the comprehensive factors of these three aspects, the regional quotas were determined. Therefore, we added other factors related to people and regions (see Section 3.1.2, for details) into the feature system. Subsequently, we chose MSE as the loss function, Adam as the gradient descent algorithm, and trained BP to obtain the final learning rate (0.003), number of iterations (5000), and other parameters (Table 11).

The total loss rate of the new model was 0.00089, and the results of the comparison between the training set and historical quotas are shown in Table 8 (new model in Table 8). The test results are displayed in Table 9 (new model in Table 9). After adding up the historical quotas in pilot regions and the estimated quotas in the nonpilot regions, the national quota was obtained (new model in Table 10).

Table 11: Parameters of the new model.

Layer	Function	Activation
Import-hidden	(21, 19)	ReLU
Hidden-export	(19, 1)	$f(x) = x$

While using the comprehensive feature system, the new model displayed a lower loss rate than that obtained with the benchmark model, and the calculated national quota was closer to the CO_2 emissions reported by China. These results indicated that the feature system has a certain degree of rationality and accuracy. Similarly, the calculation results of the model presented a trend, where the amount of national carbon quotas initially increased and later began to decrease.

6. Conclusions and Future Work

The whole world is expecting China to lead the economic recovery and green development after the global epidemic. It also expects China's 14th Five-Year Plan to become the Guide for green recovery. In the same year, China established a unified national carbon emissions market. This represents not only China's further exploration of the carbon emissions trading system to achieve green development but also one of the important tools for China to achieve two low-carbon goals. In addition, quota allocation is an important factor that determines the functionality of Chinese carbon market. In order to calculate other regional quotas, we trained a BP benchmark model. For this purpose, we considered historical quota data of China's 7-carbon emissions trading market pilot regions from 2014 to 2017 and selected suitable features that fit China's "bottom-up" total control method. Later, we built a feature system that included human, corporate, and regional factors, retrained the model, and recalculated quotas for other regions. The results are presented herein. First, both, the benchmark model results and the results obtained using the comprehensive feature system showed that within the sample interval, the amount of China's national carbon quotas displayed an initial increase to later decrease. Second, the model trained with the characteristic data of the feature system built in the present research displayed a lower loss rate as compared with the benchmark model. These results demonstrated that the feature system proposed in this paper fits not only the actual situation of

China's CO_2 emissions and quotas but also that the framework of the system is reasonable and accurate. Third, the feature system and training model proposed in the present article combined with the original "bottom-up" total control and post adjustment method can be used by Chinese CO_2 emission decision makers to obtain advanced predictions. We have provided the content of China's carbon emissions trading quota system, which can promote the operation of China's carbon emissions market, encourage participants in market transactions to reduce emissions, and accelerate China's low-carbon development.

Of course, we also admit that, in the future, the feature system and model proposed in this article can be further improved and perfected as follows. First, the indicators related to enterprises in the currently constructed feature system are substitute indicators because the specific transaction data and enterprise-related data of China's carbon emission market have not been unified and officially announced. Therefore, once the data is available, this part of the indicators will increase or decrease. Second, at present, China's national unified carbon emissions trading market has just started, and the main participants are enterprises, with less individual participation. At the same time, the central government has not issued a policy about people's low-carbon life. Therefore, when China's emission reduction program enters the critical stage in the future, the features related to people will increase or decrease. Third, the current BP neural network model has only three layers. In the future, with the improvement of feature data, a certain number of hidden layers may be further increased to train a model with low loss rate and stronger generalization ability.

Data Availability

Regional features (except that for Shenzhen) were obtained from the National Bureau of Statistics and China Statistical Yearbook (https://data.cnki.net/yearbook/Single/N2019110002). Features of Shenzhen were obtained from the Shenzhen Statistical Yearbook (https://data.cnki.net/yearbook/Single/N2020030065). The historical quotas data came from the 2014–2018 Beijing Carbon Market Annual Report (https://cbeex.com.cn/article/xxfw/xz/bjtscndhq/).

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Y.Q. designed the model and the computational framework, analyzed the data, and wrote the manuscript. W.P. conducted empirical research. R.Y. collected the relevant literatures. G.P. collected the data. All authors discussed the results and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This research was supported by the National Natural Science Foundations of China (71763010, 71463023, and 71803038) and Young Excellent Academic Talent Support Program of Jiangxi University of Finance and Economics.

References

[1] T. Balamurugan, L. Karunamoorthy, N. Arunkumar, and D. Santhosh, "Optimization of inventory routing problem to minimize carbon dioxide emission," *International Journal of Simulation Modelling*, vol. 17, no. 1, pp. 42–54, 2018.

[2] G. Nastase and A. Serban, "Experimental study on CO_2 capture in a residential space," *Environmental Engineering and Management Journal*, vol. 18, no. 5, pp. 1001–1011, 2019.

[3] J. Du, F. Qiao, and L. Yu, "Temporal characteristics and forecasting of PM2.5 concentration based on historical data in Houston, USA," *Resources, Conservation and Recycling*, vol. 147, pp. 145–156, 2019.

[4] L. Tang, H. Wang, L. Li, K. Yang, and Z. Mi, "Quantitative models in emission trading system research: a literature review," *Renewable and Sustainable Energy Reviews*, vol. 132, p. 110052, 2020.

[5] Y. Zhou, J. Jiang, B. Ye, Y. Zhang, and J. Yan, "Addressing climate change through a market mechanism: a comparative study of the pilot emission trading schemes in China," *Environmental Geochemistry and Health*, vol. 42, no. 3, pp. 745–767, 2020.

[6] A. Rose, B. Stevens, J. Edmonds et al., "International equity and differentiation in global warming policy," *Environmental and Resource Economics*, vol. 12, no. 1, pp. 25–51, 1998.

[7] S. Kverndokk, "Tradeable CO_2 emission permits: initial distribution as a justice problem," *Memorandum*, vol. 4, 1992.

[8] W. Y. Chen and Z. X. Wu, "Carbon emission permit allocation and trading," *Journal of Environmental Sciences (IOS Press)*, vol. 11, no. 4, p. 468, 1999.

[9] S. Kverndokk, E. Nævdal, and L. Nøstbakken, "The trade-off between intra- and intergenerational equity in climate policy," *European Economic Review*, vol. 69, pp. 40–58, 2014.

[10] C. O. Criado and J. M. Grether, "Convergence in per capita CO_2, emissions: a robust distributional approach," *Resource & Energy Economics*, vol. 33, no. 3, pp. 637–665, 2010.

[11] J. W. Park, C. U. Kim, and W. Isard, "Permit allocation in emissions trading using the Boltzmann distribution," *Physica A Statistical Mechanics & Its Applications*, vol. 391, no. 20, pp. 4883–4890, 2011.

[12] M. Ghiyasi, "Emission utilization permission based on environmental efficiency analysis," *Environmental Science and Pollution Research*, vol. 26, no. 21, p. 21295, 2019.

[13] H. H. Wang, H. C. Liu, X. J. He, and W. H. Zeng, "Allocation of carbon emissions right based on the intergenerational equity," *China Environmental Science*, vol. 36, no. 6, pp. 1895–1904, 2016.

[14] W. J. Yi, L. L. Zou, and J. Guo, "How can China reach its CO_2 intensity reduction targets by 2020? A regional allocation based on equity and development," *Energy Policy*, vol. 39, no. 5, pp. 2407–2415, 2011.

[15] Q. Dai, Y. Li, Q. Xie, and L. Liang, "Allocating tradable emissions permits based on the proportional allocation concept to achieve a low-carbon economy," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–8, 2014.

[16] Y. J. Hu, X. Y. Li, and B. J. Tang, "Assessing the operational performance and maturity of the carbon trading pilot

program: the case study of Beijing's carbon market," *Journal of Cleaner Production*, vol. 161, no. 10, pp. 1263–1274, 2017.
[17] J. Zhu, H. Sun, N. Liu, D. Zhou, and F. Taghizadeh-Hesary, "Measuring carbon market transaction efficiency in the power industry: an entropy-weighted TOPSIS approach," *Entropy*, vol. 22, no. 9, pp. 973–984, 2020.
[18] T. Wang, X. Wang, Y. Gong et al., "Initial allocation of carbon emission permits in power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 2, pp. 239–247, 2017.
[19] Y. Wang, H. Zhao, F. Duan, and Y. Wang, "Initial provincial allocation and equity evaluation of China's carbon emission rights-based on the improved TOPSIS method," *Sustainability*, vol. 10, no. 4, pp. 982–996, 2018.
[20] J. K. Song, L. L. Liang, and D. P. Niu, "Allocation of carbon emission permits among provinces in China:based on environmental FCAM," *Technology Economics*, vol. 36, no. 10, pp. 100–106, 2017.
[21] Q. W. Yu and F. P. Wu, "Bi-level planning model of provincial carbon emission rights allocation from the perspective of equity and efficiency," *Soft Science*, vol. 32, no. 4, pp. 72–76, 2018.
[22] H. Q. Qian, L. B. Wu, and F. Z. Ren, "From "spurring a willing horse" to efficiency driven: a study of China's regional CO_2 emission permit allocation," *Economic Research Journal*, vol. 54, no. 3, pp. 86–102, 2019.
[23] J. Wu, Y. Fan, Y. Xia, and J. Y. Liu, "Impacts of initial quota allocation on regional macro-economy and industry competitiveness," *Management Review*, vol. 27, no. 12, pp. 18–26, 2015.
[24] S. H. Zeng and Y. Xu, "Research on China's provincial carbon emission reduction allocation efficiency based on zero-sum DEA model," *Modernization of Management*, vol. 34, no. 5, pp. 63–65, 2014.
[25] H. Jiang, X. Shao, X. Zhang, and J. Bao, "A study of the allocation of carbon emission permits among the provinces of China based on fairness and efficiency," *Sustainability*, vol. 9, no. 11, pp. 2122–2134, 2017.
[26] W. Pan and W. L. Pan, "Research on the allocation of China's provincial carbon emission rights based on energy efficiency," *Soft Science*, vol. 32, no. 6, pp. 45–48, 2018.
[27] W. Guo, T. Sun, and H. Dai, "Efficiency allocation of provincial carbon reduction target in China's "13·5" period: based on zero-sum-gains SBM model," *Sustainability*, vol. 9, no. 2, pp. 167–186, 2017.
[28] B. Ye, J. Jiang, L. Miao, and D. Xie, "Interprovincial allocation of China's national carbon emission allowance: an uncertainty analysis based on Monte-Carlo simulations," *Climate Policy*, vol. 17, no. 4, pp. 401–422, 2017.
[29] H. Alishiri, A. Taklif, H. Amadeh et al., "Efficient allocation of CO_2 emissions in selected OPEC member based on zero sum gains (ZSG-DEA) data envelopment analysis model," *Quarterly Journal of Applied Theories of Economics*, vol. 5, no. 1, pp. 213–236, 2018.
[30] Y. Huang and Y. Zhang, "Energy use and carbon emissions efficiency study of Chinese regions based on price factor," *Polish Journal of Environmental Studies*, vol. 27, no. 5, pp. 2059–2069, 2018.
[31] Y. W. Qi, "Decoupling effect and gravity center trajectory of regional economic growth and carbon emissions in China," *Modern Finance & Economics*, vol. 38, no. 5, pp. 17–29, 2018.
[32] The European Union, "EC guidance document no1 on the harmonized free allocation methodology for the EU-ETS post 2012," in *General Guidance to the Allocation Methodology*, 2011, https://ec.europa.eu/clima/policies/etshttp://ccap.org/assets/Tomas-Wyns_CCAP_Ch.pdf.
[33] F. Cucchiella, I. D'Adamo, M. Gastaldi, and S. C. Lenny Koh, "Assessment of GHG emissions in Europe: future estimates and policy implications," *Environmental Engineering and Management Journal*, vol. 19, no. 1, pp. 131–142, 2020.
[34] G. Liobikienė and M. Butkus, "The European Union possibilities to achieve targets of Europe 2020 and Paris agreement climate policy," *Renewable Energy*, vol. 106, pp. 298–309, 2017.
[35] I. Perissi, S. Falsini, U. Bardi et al., "Potential European emissions trajectories within the global carbon budget," *Sustainability*, vol. 10, no. 11, pp. 4225–4239, 2018.
[36] J. B. Kim, "Implementation of artificial intelligence system and traditional system: a comparative study," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 135–146, 2019.
[37] V. Laparra, A. Berardino, J. Ballé, and E. P. Simoncelli, "Perceptually optimized image rendering," *Journal of the Optical Society of America A*, vol. 34, no. 9, pp. 1511–1525, 2017.

Hindawi
Complexity
Volume 2021, Article ID 6633906, 12 pages
https://doi.org/10.1155/2021/6633906

Research Article

Applying a Probabilistic Network Method to Solve Business-Related Few-Shot Classification Problems

Lang Wu [1] **and Menggang Li** [2]

[1]*School of Applied Science, Beijing Information Science and Technology University, Beijing, China*
[2]*School of Economics and Management, Beijing Jiaotong University, Beijing, China*

Correspondence should be addressed to Lang Wu; wulang_0306@163.com

Received 26 November 2020; Revised 9 December 2020; Accepted 11 December 2020; Published 5 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

It can be challenging to learn algorithms due to the research of business-related few-shot classification problems. Therefore, in this paper, we evaluate the classification of few-shot learning in the commercial field. To accurately identify the categories of few-shot learning problems, we proposed a probabilistic network (PN) method based on few-shot and one-shot learning problems. The enhancement of the original data was followed by the subsequent development of the PN method based on feature extraction, category comparison, and loss function analysis. The effectiveness of the method was validated using two examples (absenteeism at work and Las Vegas Strip hotels). Experimental results demonstrate the ability of the PN method to effectively identify the categories of commercial few-shot learning problems. Therefore, the proposed method can be applied to business-related few-shot classification problems.

1. Introduction

The artificial intelligence witnesses rapid development and yields successes in many fields due to recent explosive growth of data volume and increasing computer processing power [1–5]. With minimal or few pictures, humans can recognize objects based on previous experience and knowledge. It is attributed to the ability of learning from prior knowledge accumulated through years of experience. For example, users can quickly learn how to use smart phones due to their previous knowledge of using Nokia cellphones. Therefore, skills should be trained early by constructing an intelligent system to master multiple skills and to adapt to various environments. And new tasks should be learned based on previous experience.

Few-shot learning can be employed for this purpose. It refers to the ability of generalizing a model with a limited number of labeled samples, particularly "drawing inferences from one example." As an important component in intelligent systems, few-shot learning experienced an extensive development.

Research studies on few-shot learning began in the 1990s [6–8], coinciding with the application of sparse representation. Early studies focused on the exploratory work and proved to be successful. Several influential studies were published from 2003 to 2015, including the work on Bayesian programming learning [9–13]. Li and Fergus et al. proposed the concept of one-shot learning for the first time and demonstrated its application in classifications based on the Bayesian framework [9, 10]. The major characteristic of the Bayesian learning stage is that it can handle small samples and be well integrated into the prior knowledge because the methods adopted by the model are often based on Bayesian theory. However, one problem exists, that is, the generality of the model is often insufficient. In other words, the model at this stage is often designed for a specific problem. When the problem changes, consequently the model is no longer valid. Lake [12, 13] combined the Bayesian framework with prior knowledge to validate concept learning, proving that machine can use background knowledge through complete probability estimation based on few-shot learning. Such methods are highly dependent on

selecting prior knowledge of human. The methods adopted by the Bayesian learning stage are often based on Bayesian theory because they can deal with the few-shot learning problem and to integrate prior knowledge. However, the model generality at this stage is typically insufficient. The model is usually designed for a specific problem. When the problem changes, the model is no longer valid.

Since 2015, most of the research studies on few-shot learning have focused on neural networks [14–18]. For example, the same model can be used for cat and dog recognition and face recognition under the small sample. Except for the adjusted super parameters, the model itself is unchanged, resulting in the analysis and modeling of few-shot learning problems from various angles. Currently, few-shot learning can be divided into the model-, optimization-, and metric-based methods.

(i) The model-based approach [19, 20] understands the task through a model and stores the knowledge acquired. When decisions are required, they can be made by learning effective models. As early as 2001, the memory-based neural network method was proved to be applicable to meta-learning [19], whereby the bias and output were adjusted by updating weights and learning to quickly cache expressions into the memory, respectively. The authors used long short-term memory (LSTM) and other recurrent neural network (RNN) to treat the model data as a sequence for training and inputted new class samples for classification during testing. Meta networks [20] were designed for the rapid generalization between tasks by introducing the concept of fast weight. In particular, the gradient descent algorithm is used to optimize the weights in the neural networks, which is typically a slow process. It can be accelerated with the use of an additional neural network to predict the weights, denoted as fast weight. And the weight optimized by the general gradient descent algorithm is denoted as the slow weight. In the meta-network, the lost gradient is provided as meta-information to learn the fast weight model. The slow and fast weights are subsequently combined for the final output.

(ii) The ability that fulfills the few-shot learning is employed in the optimization-based few-shot learning models. For example, Finn et al. proposed the model-agnostic meta-learning (MAML) in 2017 [21]. MAML is a general optimization algorithm that expands the differential calculation process through the computational diagrams of the gradient descent method and learns a model that includes tasks but not samples. Nichol et al. proposed Reptile [22], a simple meta-learning optimization algorithm that is similar to MAML. For example, Reptile and MAML are gradient-based meta-optimizations and are model-independent. The optimizer performs a multistep gradient descent algorithm on each training task and updates the model with the results of the last step.

(iii) Metric learning tasks learn a function that measures the distance between different objects. Measurement learning can be combined with distance and similarity-based methods, such as K-mean clustering [23], the K-nearest neighbor method [24], support vector machine [25], and other algorithms that require a given measure to reflect the important relationship among data. In the Siamese network, two networks with the same parameters are used to extract features of the two samples. Then, the extracted features are inputted into the discriminator to determine whether the two samples belong to the same object class [26]. The matching network builds encoders for the supporting and query sets, and the output of the final classifier is the weighted sum of the predicted values between the supporting and query set samples [27]. The prototype network maps the sample data in each category to a given space and extracts their "mean" and Euclidean distance to represent the class prototype and distance measurement, respectively. Thus, the training data and class prototype exhibit the closest distance compared to other prototypes [28].

The above research studies on few-shot learning generally focus on images, while studies on the practical application of few-shot learning problems are limited. Thus, in the current paper, the problem of few shot learning in the commercial field is investigated because it remains a challenge. The work is performed in terms of two aspects. On the one hand, we proposed the probabilistic network (PN) method for the business field by adopting data enhancement to process the original dataset so as to improve the sensitivity of the model to business data. On the other hand, we ensure the effective application of the PN and experimental results of absenteeism at work. Las Vegas Strip hotels can valid that the PN method could identify the categories of commercial few-shot learning problems.

The following parts are structured as follows. The applications of the PN method for few-shot and one-shot learning are described in Section 2. The experimental results are presented in Section 3. Section 4 discusses the significance of business-related few-shot classification problems. The conclusion is drawn in Section 5.

2. Methodology

Few-shot learning is an integral component for research on the exploration of intelligent systems. Current smart systems are often based on large amounts of data analysis, while few-shot learning exhibits the learning ability of extracting inferences from a single example.

Unmarked samples can be encountered extensively in real-life applications, particularly in business field. The inaccurate prediction or data classification will directly affect the business development. However, traditional forecasting methods cannot always accurately foresee business problems. Therefore, more effective indicators and models must be built in order to improve the forecasting accuracy,

understand the development trends of business problems, and promote economic growth.

2.1. Data Augmentation. Data enhancement allows a limited dataset to generate more data, increases the number and diversity of training samples (noise data), and improves the robustness of the model. Therefore, in order to effectively solve the problem of few-shot learning prediction for business applications, we expand the original dataset to increase the difference between the data, thus improving the final prediction. In particular, we collect and learn the mathematical distribution of real sample data X and add noise data λ to generate new sample data X^*:

$$X^* = (X, \lambda)^T (X, \lambda). \tag{1}$$

New sample data X^* can be described by n samples as follows:

$$X^* = (x_1, x_2, \ldots, x_n), \tag{2}$$

thus reflecting the diversity of the data.

2.2. Problem Definition. In the current paper, we consider the one- and few-shot classifier learning tasks. Each task is made up of four datasets, namely, the training, test, support, and query sets. The support and query sets share the same tag space. The goal of the task is to decide which class the query belongs to.

2.2.1. Training. The episode training strategy is adopted in the training phase of the proposed method. This strategy breaks the training process into multiple episodes and subsequently performs the classification on the task for each episode. The few-shot problem training set contains several categories, each of which has many samples. In the training stage, Q categories will be randomly selected from the training set, with K samples in each category (total $Q * K$ data) to construct a subtract task. The subtract task will then be inputted as the support set (SS) of the model:

$$\mathrm{SS} = (x_i, y_i), \quad i = 1, 2, \ldots, Q * K, y_i \in 1, 2, \ldots, Q. \tag{3}$$

A sample of the remaining data from the Q classes is then extracted as the model's query set (QS):

$$\mathrm{QS} = (x_i, y_i), \quad i = 1, 2, \ldots, Q * (m - K), y_i \in 1, 2, \ldots, Q, \tag{4}$$

where m is the total number of sample of Q classes. Table 1 reports the rules used for the division of X^* to determine SS and QS.

In Table 1, random sample (SS, N) takes a random sample N without returning it back to SS; X^*/SS denotes the elements that are in collection X^* but not in collection SS.

In the business field, a Q-way K-shot problem denotes the model used to learn how to distinguish the Q categories from the $Q * K$ data and is considered as a business task. Thus, each episode will have distinct business tasks throughout the entire training process. We denote the entire training process as T_{train}, while each trained business task is defined as $T_{\mathrm{train}}(\tau)$, $\tau = 1, 2, \ldots, E$, where E is the number of business tasks. The business tasks satisfy $T_{\mathrm{train}}(\tau) \in T_{\mathrm{train}}$, $\tau = 1, 2, \ldots, E$. $T_{\mathrm{train}}(\tau)$ can be described as follows:

$$\begin{aligned} T_{\mathrm{train}}(\tau) &= \mathrm{SS}_\tau \bigcup \mathrm{QS}_\tau \\ &= \left\{\left(x_1^\tau, y_i^\tau\right), \left(x_2^\tau, y_2^\tau\right), \ldots, \left(x_{QK}^\tau, y_{QK}^\tau\right)\right\} \bigcup \left\{\left(x_1^\tau, y_i^\tau\right), \left(x_2^\tau, y_2^\tau\right), \ldots, \left(x_{Q(m-K)}^\tau, y_{Q(m-K)}^\tau\right)\right\}. \end{aligned} \tag{5}$$

Next, we employ the training set to constantly adjust the parameters and determine the final training model.

2.2.2. Test. The evaluation stage adopts the same framework as the training stage, but with data taken from the evaluation dataset. The test process is defined as $T_{\mathrm{test}}(\tau)$, $\tau = 1, 2, \ldots, E$ and also contains the test set (TS), which described as follows:

$$\mathrm{TS} = (x_i, y_i), \quad i = 1, 2, \ldots, Q * K, y_i \in 1, 2, \ldots, Q. \tag{6}$$

The datasets are then used to evaluate the generalization ability of the final model.

2.2.3. Task. The category labels for the sample in the query set are determined from the sample in the support set. Hence, the training and testing processes make up the entire learning task. For this task, we trained the model, adjusted the parameters, and predicted the data in order to solve the one- and few-shot problems. The learning task is defined as T_τ and satisfies the following formula:

$$\begin{aligned} &T_\tau \coloneqq \left(T_{\mathrm{train}}(\tau) \bigcup T_{\mathrm{test}}(\tau), P\left(Q|x_i\right), \left(x_i, y_i\right)\right), \\ &\tau = 1, 2, \ldots, E, \\ &i = 1, 2, \ldots, n, \\ &y_i \in \{1, 2, \ldots, q\}, \\ &Q \le q, \end{aligned} \tag{7}$$

where the dataset contains class q. We then divided the problem into two categories depending on category Q:

Problem 1. Few-shot business problem: it is defined by $k > 1$ and $\mathrm{SS} = QK$. The sample number of each class in the support set is greater than one, and the number of query sets is unconstrained. This problem is denoted as the few-shot problem of finance.

Table 1: Sampling rules of the dataset.

Rule: data sampling
Input: dataset X^*
Output: **support** set (SS), **query set** (QS)
1: $V \longleftarrow$ Random Sample $(\{1, 2, \ldots, n\}, Q)$
2: for i in $\{1, 2, \ldots, Q\}$:
3: $\quad SS_i \longleftarrow$ Random Sample $(X^*_{V_i}, K)$
4: $\quad QS_i \longleftarrow$ Random Sample $(X^*_{V_i}/SS_i, t(m-K))$
5: return SS, QS

Problem 2. One-shot business problem: this is defined by $k = 1$, where the sample numbers of each class is one in the support set, and the number of query sets is unconstrained. This problem is denoted as the one-shot problem of finance.

2.3. PN Category Method for the Few-Shot Problem. In this section, the classification of the few-shot problem is described.

2.3.1. Feature Embedding. The neural network is widely used to construct new feature spaces, as it is considered as a successful technique in feature embedding based on their strong learning ability. Convolutional neural networks are locally connected neural networks with advantages in mapping, classification, prediction, and learning speed. The feature embedding based on the convolutional neural network is performed under the following steps:

*S*1 (input layer): input sample x_i, and rewrite it as $\hat{x}_i$;

*S*2 (convolution layer): data $\hat{x}_i$ are scanned using the deep convolutional neural network $g(\hat{x}_i; \omega, b)$, where convolution kernel $\omega \in R$ is a learnable weight vector and $b \in R$ is the learnable bias, to obtain a feature map;

*S*3 (batch norm layer): batch norm standardization is adopted to ensure the equal distribution of the layer inputs. Thus, the network has the following form:

$$g(\hat{x}_i, t(\hat{x}_i, tEn(\hat{x}_i) qvarh(\hat{x}_i)) n; q\omega h_{,b}), \tag{8}$$

where $E(\hat{x}_i)$ and $\mathrm{var}(\hat{x}_i)$ are the mean and variance of word vectors $\hat{x}_i$, respectively.

*S*4 (rectified linear unit (ReLU) nonlinearity layer): in order to avoid linearization, the following activation function is used:

$$f(g(\hat{x}_i; \omega, b)) = f_g(\hat{x}_i), \tag{9}$$

where f is the ReLU activation function.

The parameters of the feature extraction network are maintained consistent for each task across the support and query sets. Figure 1 summarizes the feature extraction framework.

The convolutional layer aims to extract local features of the feature matrix, in order to reduce the computational complexity of the network and to obtain more representative features. Moreover, the batch norm has a fundamental influence on the training process by smoothing the solution space of the optimization problems. This smoothness ensures that the gradient is both predictable and stable, allowing for a wider range of learning rates and a faster network convergence. The batch norm is applied to the sensitive areas of the activation function, corresponding to the frontal section of the function in this paper. Training difficulties often occur during the training of deep networks, which is attributed to variations in the output data of the upper layer network, following each parameter iteration and update. Such variations are denoted as internal covariate shifts and complicate the network learning of the next layer. In order to reduce the internal covariate shift, each layer of the neural network is normalized.

2.3.2. Category Contrast. Once the feature map is obtained by randomly extracted samples from the support set $\hat{x}_i$and query set $\hat{x}_j$, a key problem arises related to the classes learning from a small set. In our method, the final decision is made by combining the query set sample with the information for each category via the category determination system.

The network processing of two feature maps $f_g(\hat{x}_i)$ and $f_g(\hat{x}_j)$ is compared, and the corresponding scores are calculated. Each score denotes the probability of each classification of the query set sample $\hat{x}_j$. It is noted that for $Q > 1$, we determine the Q scores. The feature map outputs are then summed element-by-element. Furthermore, the target function can be expressed as follows:

$$\begin{aligned} &\max P(q|\hat{x}_j; \varphi(\mathrm{combinate}(f_g(\hat{x}_i), f_g(\hat{x}_j)), \\ &s.t. \left|\varphi(\mathrm{combinate}(f_g(\hat{x}_i), f_g(\hat{x}_j))\right| < \varepsilon, \hat{x}_i, \hat{x}_j \text{ belong to the same category}, \\ &\left|\varphi(\mathrm{combinate}(f_g(\hat{x}_i), f_g(\hat{x}_j))\right| > \varepsilon, \text{otherwise}, \end{aligned} \tag{10}$$

where $q = 1, 2, \ldots, Q$, $0 < \varepsilon < 1$. Probability P is determined using the probability network (PN) method, described in the following, where $\varphi(x)$ represents the network structure.

(1) Convolution layer: the convolution module contains a convolutional layer, a batch normalization layer, and a maximum pooling layer

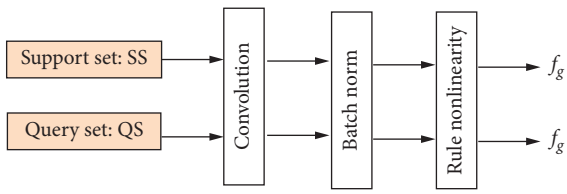


Figure 1: Feature embedding framework of the proposed method.

(2) Fully connected layer 1: it is a layer with an ReLU activation function and a layer that can combine all the local features into global features to calculate the final score for each category

(3) Fully connected layer 2: it is a linear layer with a Sigmod activation function

The neuron is connected by the fully connected layer to the neuron in the next layer. The fully connected layer can act as a "classifier" for the whole neural network, mapping the learned features to the sample marker space. A nonlinear activation function is introduced to the neural network to determine the categories of the query set samples.

At an input of 0, the ReLU activation function equals 0. But when output values equal the input value, the input is greater than 0. It increases the network sparsity, reduces the amount of computation, and enhances the convergence speed. What is more, it is commonly applied by current mainstream networks. When inputs are greater than 0, the slope is constant, and the gradient disappearance can be reduced by the ReLU.

The Sigmod function maps the input between [0, 1]. When the absolute value of the input is large, the output is close to 0 or 1 and output changes are minimal. The Sigmod function can be considered as a sample layer in the output data, whereby the greater output value means larger probability of sample labels. On the contrary, with smaller output data value, the sample label features less probability. Thus, as the input value increases, the effect of increasing probability is gradually diminished and vice versa, highlighting the benefits of nonlinear mapping.

The activation functions ensure that the classification probabilities of the query set sample are between 0 and 1. In theory, the sample category is determined based on the maximum output results. It effectively considers which class properties may be most relevant to the category of unknown sample. Figure 2 summarizes the category contrast.

2.3.3. Loss Function. Applying the sigmoid function as the neuron activation function, we employ the l_1 loss function to replace the variance cost function in order to avoid a slow training process. The l_1 loss function is frequently used in classification problems and represents the difference between the actual sample tag and the predicted probability. More specifically, the smaller the loss function value is, the closer the actual sample tag to the predicted probability will be.

The number of nodes in the last output layer is generally equal to the number of targets of the classification task. Let $N(\mathrm{QK})$ denote the final number of nodes; then, the neural network is able to determine an n-dimensional array as the output result for each sample, with each array dimension corresponding to a category. In the most ideal case, if a sample belongs to the class, then the output value of the output node corresponding to the category should be 1, while that of other nodes should be 0, namely [0,0,1,0, . . ., 0]. This array is taken as the sample label and is the most expected output result of the neural network. Thus, the l_1 loss function between the actual sample tag and the predicted model can be expressed as follows:

$$L_{\mathrm{loss}} = \frac{1}{M}\sum_{\tau=1}^{M}\left|P\left(q|\hat{x}_j;\varphi\left(\mathrm{combinate}\left(f_g\left(\hat{x}_i\right),f_g\left(\hat{x}_j\right)\right)\right)-\delta\left(\hat{x}_i,\hat{x}_j\right)\right)\right|,$$
$$\delta=\begin{cases}1, & \hat{x}_i,\hat{x}_j \text{ belong to the same category},\\ 0, & \text{otherwise}.\end{cases} \tag{11}$$

In which, M expresses the number of subtasks to learn a model. The norm in the loss function ensures a stable gradient, irrespective of the input value. Hence, this is a relatively robust solution that avoids the gradient explosion.

2.3.4. Training Strategy. The effective feature learning for a relatively small amount of labeled data proves to be a complicated task. These rare categories need to be generalized without additional training due to the associated costs and long cycles. In order to enhance the classifier generalization capability, few-shot learning generally employs episodic training. Thus, a large number of subtasks T_τ (or historical tasks) similar to the target task are taken to learn a model. A reasonable initial model value is then obtained by acting on the target task, such that the model can rapidly adapt with just a small amount of data for the target task. However, the model needs to combine previous experience with the information learned by the few-shot approach of the current new task. In addition, overfitting must be avoided. We take the entire classification task as T, which can be described in terms of subtasks T_τ:

$$T=\sum_{\tau=1}^{M}T_\tau, \tag{12}$$

where T_τ satisfies equation (7).

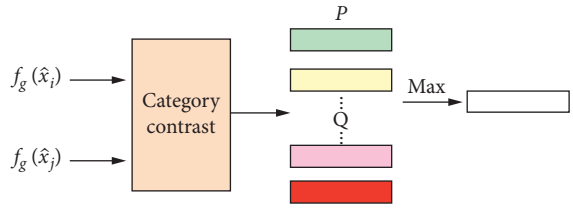


Figure 2: Schematic diagram of the category contrast process.

The model predicts the sample labels for each subtask in the query set for a given support set, where the purpose of training is to minimize the prediction error. This process can be considered as meta learning as the training process explicitly uses a given support set for learning in order to minimize the query set error. The l_1 loss function between the actual sample tag and the predicted model is adopted in this paper. And the smaller loss function can produce a robust model, which is in the following expression:

$$\min L_{\text{loss}}\varphi\left(\text{combinate}\left(f_g(\hat{x}_i), f_g\left(\hat{x}_j\right)\right), \delta\left(\hat{x}_i, \hat{x}_j\right)\right). \quad (13)$$

The training process described in equation (5) will output a set of parameters. When faced with new data categories, the model will exhibit a strong performance under these parameters. The sampled tasks are distinct when training is performed each time. Hence, the training includes numerous category combinations. It enables the model to learn the common components of different tasks, extract important features, and perform category comparisons. The models learned through this learning mechanism can effectively classify new tasks. Figure 3 summarizes the training process.

2.4. Category Method of the One-Shot Problem. For the one-shot problem, $K = 1$, namely, the query set only has one sample for each category. A convolutional neural network is adopted to accurately extract the sample features, followed by the implementation of the batch normalization layer and the ReLU activation function layer. Moreover, the parameters of the feature extraction network are consistent for each task across the support and query sets.

The network structure imitates that of the feature extractor to attain the final category. At this stage, we employ the convolution layer, batch normalization, pooling layer, fully connected ReLU, and the fully connected sigmoid nonlinearity layer. The single sample in the query set determines the final sample category by comparing the relationship between each category and assessing the category based on the output results.

3. Results

Few-shot learning consists of the training and testing stages. As the datasets used in the training and evaluation stages have no common category, episode training is typically employed in the training stage to ensure that the few-shot learning is strictly met. The training process is divided into multiple episodes by the strategy, whereby each episode requires the sampling of the task data for the task classification. The same episode training strategy is used for the testing stage. The only difference is that the data are sampled from the evaluation dataset. More specifically, the evaluation phase needs to sample the support and query sets from the evaluation dataset. The few-shot learning model needs to determine the category label of the samples in the query set based on the support set samples.

The same network and training hyperparameters are used for each task. The networks were trained in PyTorch by using the Adam optimizer with a learning rate of 0.001. Each model is trained for 100 or 200 epochs, with $Q = 3$ and $Q = 5$ per epoch. The data were obtained from the UCI Machine Learning Repository datasets (https://archive.ics.uci.edu). Based on the few-shot learning size of the business industry, we selected two datasets to construct the model datasets following definitions 1 and 2.

The dataset selection should follow two criteria. On the one hand, the data must be authentic, so the datasets come from a public database to reflect the persuasive of the results. On the other hand, it is difficult to obtain a large number of samples or marker samples in the dataset domain. Only in this way, can the value of proposed method be reflected in this paper.

The proposed method is compared with commonly used intelligent prediction methods including the *K*-nearest neighbor [29], logistic regression [30], decision tree [31], and Naive Bayes [32] algorithms. The selection of the training, validation, and test sets is consistent across methods to ensure a horizontal comparison of the differences in accuracies. However, the support, query, and test set samples are randomly selected to ensure differences in the results of each calculation.

3.1. Few-Shot Business Problem. The training set contains numerous categories, with multiple samples in each category. At the training stage, 5 categories are randomly selected from the training set, with 5 samples for each category (total of 25 samples) used to construct a support set. The model learns how to distinguish the 5 categories within the 25 samples by extracting a sample batch from the remaining class data as the query set. In addition, the proposed method also learns how to distinguish the 3 categories within the 15 samples by extracting a sample batch from the remaining class data as the query set.

The *K*-nearest neighbor, logistic regression, decision tree, and Naive Bayes algorithms adopt 5 categories to train the models and 5 categories within the 25 samples to test the models.

3.1.1. Example 1 (Absenteeism at Work). It is common to observe that employees sometimes are late for work, leave early, or are absent at work, owing to health problems and family crises. It occasionally causes absenteeism. However, such phenomenon rarely occurs in the company; thus, absenteeism datasets are small. Under such circumstances, the data mining process can be complicated, thus preventing an effective classification by traditional methods. In the

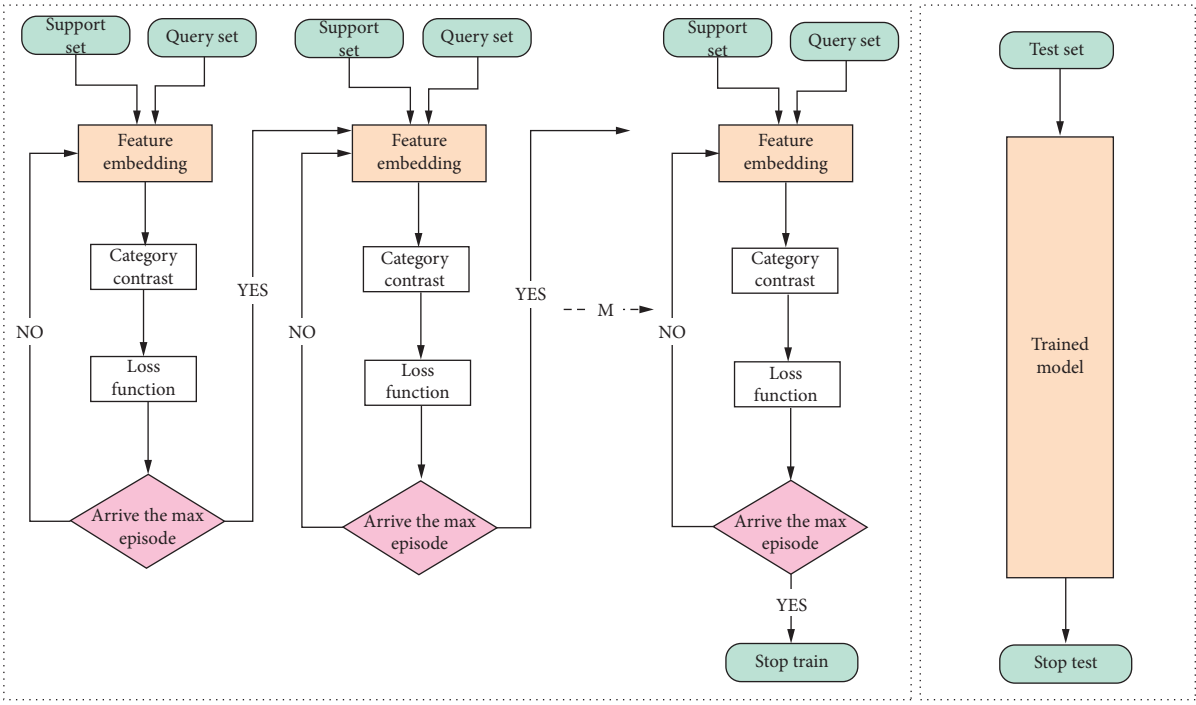


Figure 3: The application of the PN method for few-shot learning.

current paper, absenteeism records are selected from a Brazilian courier company to determine employee absenteeism. This dataset contains 740 samples and 21 variables. Variables include individual identification (ID), reason for absence, month of absence, day of the week, season, age, workload average, education, drinking status, and absenteeism time in hours.

In the simulation experiment, the convolutional neural network is employed to optimize the model and train the parameters by using Python. Table 2 indicates the accuracy of the test process. We consider both the recognition and identification accuracy. The parameter derivative typically changes to 0, preventing the parameter from updating. Figure 4 depicts the loss function changes during the training process.

The classification model proposed in this paper with other machine learning predictions (K-nearest-neighbor, logistic regression, decision tree, and Naive Bayes algorithms) is given in Table 2. The accuracy of the proposed method is higher than other machine learning prediction algorithms, which indicates our model is suitable to solve few-shot learning problems in the commercial field. In particular, our model can improve the accuracy of both the testing and classification processes.

The loss function convergence is executed with 100 iterations (Figure 4). It is close to 0.1 at iteration 100. This indicates the successful classification effect of the model as well as the ability of the proposed method to classify 3-category and 5-category problems (i.e., $Q = 3, 5$ in Problem 1).

Thus, it is demonstrated that the proposed model can not only deal with few-shot learning absenteeism effectively but also shows great advantages compared with other methods.

The working time is an important labor resource controlled by individuals. Workers achieve a given work performance by completing tasks within a limited time. With excessive absenteeism, the work will be completed with low efficiency, which can affect the performance appraisal of employees [33]. Therefore, a worker must effectively organize their working time. When the working hours are given within a reasonable range, the employees can fulfill their duties within a sufficient time frame. However, the employee will face greater time pressure if the working time is lower than the reasonable range, which consequently exerts a negative impact on work performance. Absenteeism is not only conducive to the management of employees but also affects the normal operation of the enterprise. Therefore, employee absenteeism has a far-reaching impact on business activity. It is of positive significance for employees and enterprises to effectively identify employees' absenteeism and reasonably predict the time when they are absent. It can promote the development of enterprises and bring benefits to their work performance.

3.1.2. Example 2 (Las Vegas Strip). With the development of the Internet and e-commerce, online commenting has gained wide popularity. Online user reviews serve as a vital information resource for consumers, playing a crucial role in the decision-making of potential customers. More than 80% of online consumers will refer to the comments made by other consumers before making purchasing decisions. It is believed that such information is more authentic than the information provided by sellers. Research shows that consumer reviews have a significant impact on product sales.

Table 2: Comparison of classification results on absenteeism with $k > 1$.

Method	$\mathbf{Q} = 3$	$\mathbf{Q} = 5$
K-nearest neighbor	–	42%
Logistic regressive	–	43%
Decision tree	–	45%
Bayes	–	39%
PN	94%	98%

The online ecosystem of the hotel and tourism industry is both complex and diverse. In the current paper, we focus on the lack of online reviews in this industry, taking 21 hotels in the Las Vegas Strip as an example. The dataset contains 504 records and 20 tuned features, including traveler type, tennis court, casino, free Internet, hotel name, hotel stars, no. of rooms, user continent, member years, and review month.

We follow the same classification methodology as the previous example in Section 4.1.1. Table 3 compares the accuracy of the results by using the methods. The accuracy of proposed is higher than *K*-nearest neighbor, logistic regression, decision tree, and Naive Bayes algorithms. Therefore, it is proved that the PN method is more applicable for few-shot learning.

The loss function is taken as the benchmark to determine the weight parameter minimized in equation (5). Figure 5 presents the variations in the loss function curve of the training samples, and it is close to 0.1 at iteration 200. It is observed that the PN method can effective treatment the Las Vegas Strip dataset.

Our method outperforms the traditional machine-learning in terms of accuracy (Table 3). It verifies the effectiveness of our method in dealing with problems in the business. Compared to the traditional approaches, our method obtains improved weight parameters from the training data, with stable variations in loss function (Figure 5). Thus, it is demonstrated that the proposed model can not only effectively deal with few-shot learning Las Vegas Strip but also shows great advantages compared with other methods.

Online reviews typically included reviewer information, a review rating, and review content. Consumers generally describe information related to product attributes or its performance. The review plays an important role in the purchasing decisions of potential consumers, particularly in the hospitality industry. Prior to consumption, consumers are unable to make reasonable judgments on the quality of the hotel. If a hotel is located in a city unfamiliar to consumers, online comments is of vital significance. As comments are not limited by location, users can quickly find the required information. Essentially, online reviews provide an important basis for potential consumers to make purchasing decisions and will also affect the profit of the hotel industry. Therefore, it is highly important for hotels to identify online review information. The earlier stage witnesses less the hotel online review information. It needs a small number of sample hotel online evaluation problem classification methods and accurate identification of customer evaluation. According to the classification results, the hotel can attract customers and provide better service by adjusting the business model.

3.2. One-Shot Business Problem. The training set contains numerous categories, with multiple samples in each category. During the training stage, 5 categories are randomly selected from the training set, with one sample for each category (a total of 5 samples) used to construct a support set. The model learns how to distinguish the 5 categories from the 5 samples by extracting a sample batch from the remaining data of the 5 classes as the query set. In addition, the proposed method also learns how to distinguish the 3 categories from the 3 samples.

The *K*-nearest neighbor, logistic regression, decision tree, and Naive Bayes algorithms also adopt the 5 categories to train the models.

3.2.1. Example 3 (Absenteeism at Work). In order to further verify the PN method in dealing with a one-sample problem, we continue with the problem in example 1 and take a single sample from each sample. The single sample is used to train the model, update the parameters, and determine the final classification model.

The performance of the PN method surpasses that of the other methods (Table 4). In particular, the accuracy of the PN method is 92% with $Q = 5$, exceeding the traditional approaches (the *K*-nearest neighbor, logistic regression, decision tree, and Naive Bayes algorithms) by approximately 50%. In addition, the PN loss function stabilizes, followed with 100 training iterations, which indicates the stable change in parameters (Figure 6). Therefore, the proposed PN method can effectively deal with single-sample absenteeism problems.

3.2.2. Example 4 (Las Vegas Strip). Similarly, in order to further verify the effectiveness of the proposed method for single-sample problems, we evaluate one-sample online hotel rating data. We compared the accuracy of the PN method proposed in this paper with that of traditional approaches. Table 5 reports the results, and the accuracy of proposed PN method are higher than *K*-nearest neighbor, logistic regression, decision tree, and Naive Bayes algorithms, and Figure 7 presents the loss function diagram, revealing the ability of the proposed PN method to handle the one-sample Las Vegas Strip problem.

4. Discussion

The rapid development of artificial intelligence in recent years is largely based on data development, with the increasing focus of emerging artificial intelligence put on big data. However, the data available for many real-world scenarios are limited. In this way, the acquired data can hardly be guaranteed as the massive data. Under such circumstances, conventional algorithms can hardly handle such issue, but it can be overcome through our method. Few-shot learning is able to deal with small dataset tasks. In

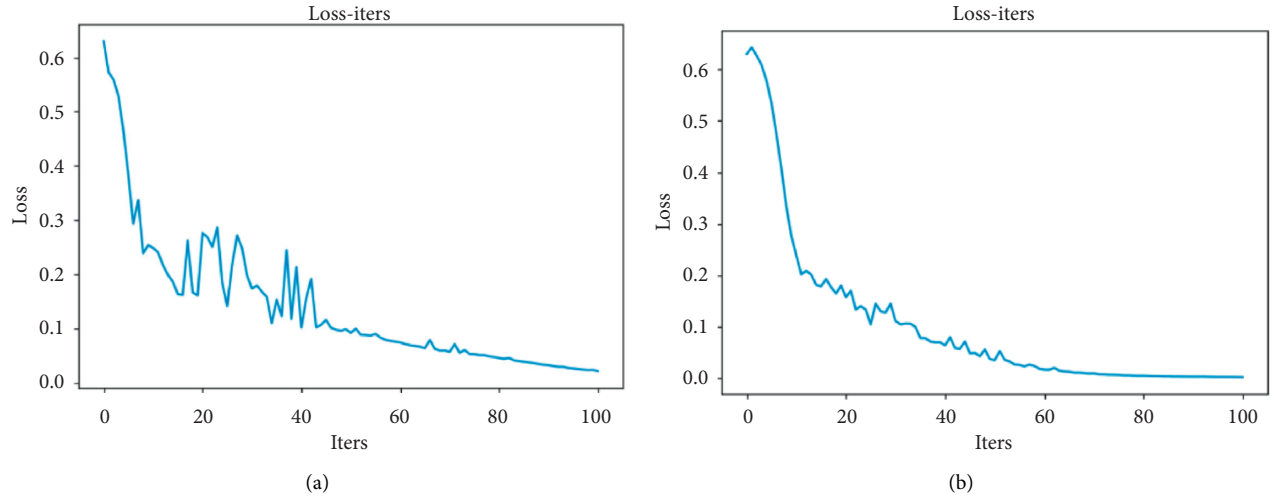


Figure 4: Variations in the loss function determined from data on absenteeism with $k > 1$. (a) $Q = 3$. (b) $Q = 5$.

Table 3: Comparison of classification results of the Las Vegas Strip dataset for $k > 1$.

Method	**Q** = 3	**Q** = 5
K-nearest neighbor	–	38%
Logistic regressive	–	39%
Decision tree	–	41%
Bayes	–	35%
PN	87%	89%

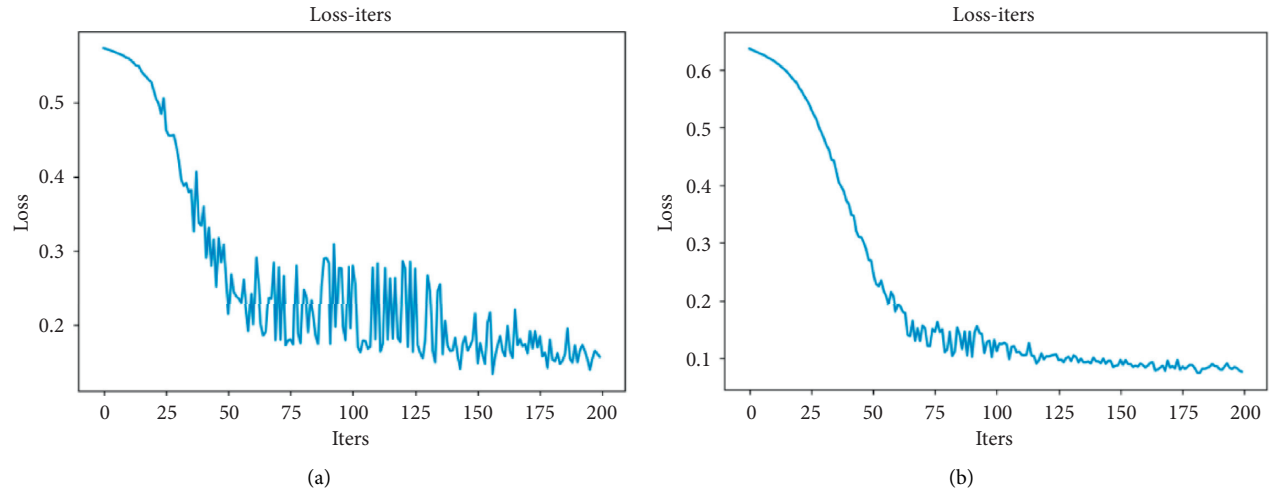


Figure 5: Variations in the loss function of the Las Vegas Strip dataset for $k > 1$. (a) $Q = 3$. (b) $Q = 5$.

Table 4: Comparison of classification results for absenteeism with $k = 1$.

Method	**Q** = 3	**Q** = 5
K-nearest neighbor	–	43%
Logistic regressive	–	43%
Decision tree	–	44%
Bayes	–	38%
PN	92%	95%

particular, it describes the task of learning from a few examples, which poses a great challenge for current machine learning algorithms. In applied statistics, few-shot learning and large samples are generally associated with smaller and larger sample sizes, respectively. In the current paper, it is used to describe a small amount of data and marked samples, a business that has failed to get established in the early stage or a large number of samples. However, accurate predictions are required to aid enterprises to expand their business and

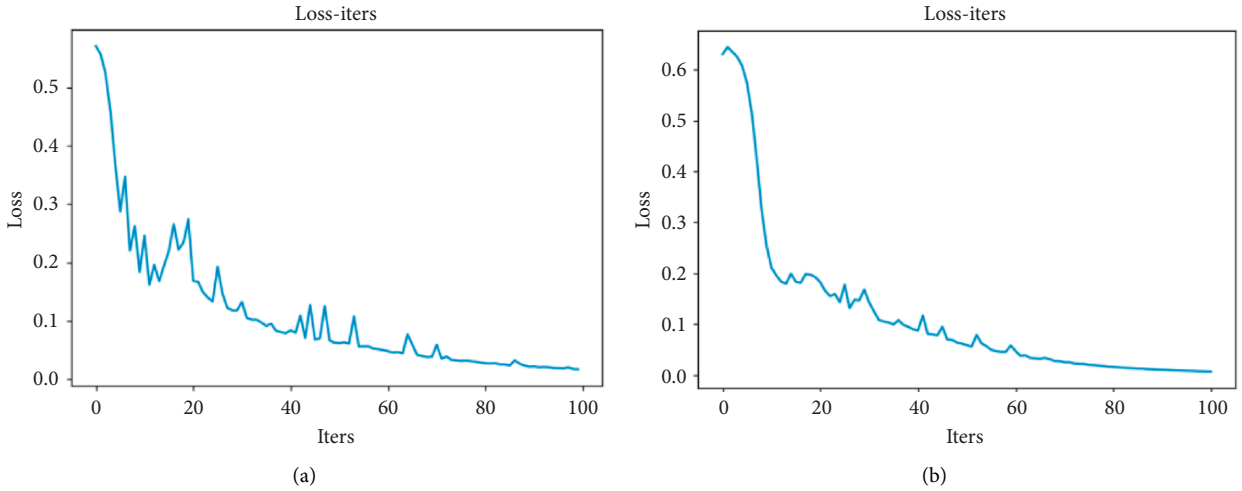


Figure 6: Variations in the loss function determined with $k = 1$. (a) $Q = 3$. (b) $Q = 5$.

Table 5: Comparison of classification results for the Las Vegas Strip dataset with $k = 1$.

Method	**Q** = 3	**Q** = 5
K-nearest neighbor	–	36%
Logistic regressive	–	38%
Decision tree	–	42%
Bayes	–	36%
PN	80%	85%

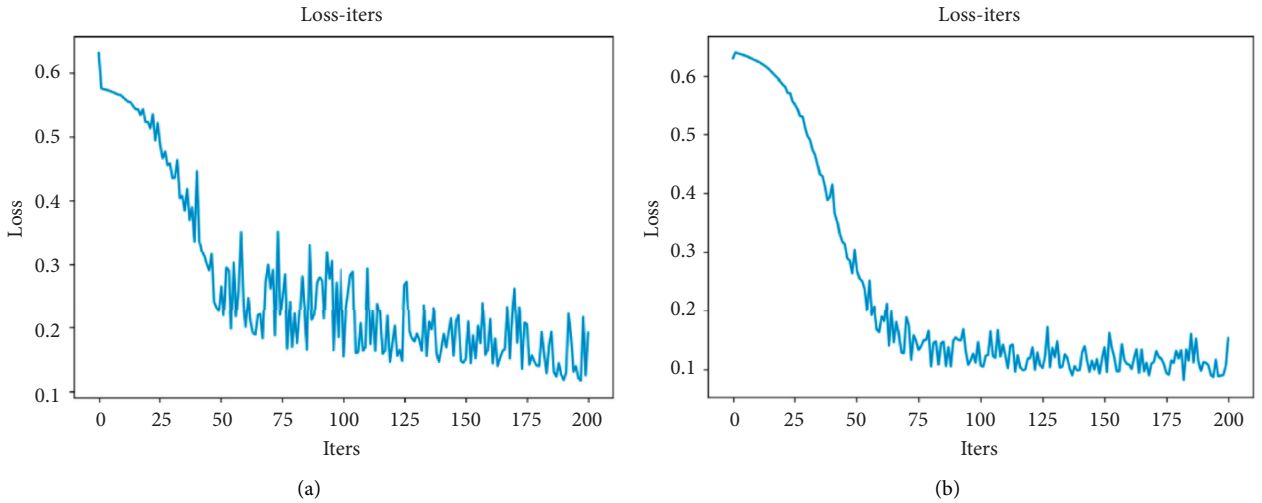


Figure 7: Variations in the loss function for the Las Vegas Strip dataset for $k = 1$. (a) $Q = 3$. (b) $Q = 5$.

enhance their competitiveness and service. The few-shot learning problem boils down to the lack of information in essence; thus, the potential data information cannot be fully mined by machine learning. To overcome this, a more intelligent technique is needed to effectively handle fewer sample tasks and achieve a higher model accuracy.

Solving the business-related few-shot problem can create new opportunities and challenges for the development of the business industry. Through the strengthened application and guidance of science technology, business enterprises can achieve sustainable and high-quality development, thus enhancing their competitiveness and maximizing enterprise trading profits.

5. Conclusion

In the current paper, the few-shot learning for business applications is investigated. In particular, we developed the PN method with feature extraction, analogy comparisons, and loss function analysis. Furthermore, two case studies are adopted on the prediction of absenteeism and online hotel reviews to validate and prove the generalization ability of the proposed method. It is demonstrated by the results that the proposed method is of high accuracy in dealing with these two problems and has the stable variation of the corresponding loss function. Thus, the method proposed in this paper can witness a successful prospect by being applied to commercial few-shot learning. However, the business-related few-shot regression problems await to be solved, which can enrich the predicted algorithm. Even though further details are still under dispute, yet it remains as the primary direction of future research.

Data Availability

The data used to support the findings of this study were obtained from the UCI machine learning repository datasets (https://archive.ics.uci.edu).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

All authors contributed equally to the manuscript and typed, read, and approved the final manuscript.

Acknowledgments

This work was supported by the foundation of Beijing Information Science and Technology University (2025034).

References

[1] Z. Zhang, Z. L. Guan, J. Zhang, and X. Xie, "A novel job-shop scheduling strategy based on particle swarm optimization and neural network," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 699–707, 2019.

[2] D. Lapkova, "Education in professional defense-possibilities of classification of training level with the help of impulse," *Journal of System and Management Sciences*, vol. 8, no. 1, pp. 23–44, 2018.

[3] J. Yoon and S. Joung, "A big data based cosmetic recommendation algorithm," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 40–52, 2020.

[4] H.-M. Afify, K.-K. Mohammed, and A.-E. Hassanien, "Multi-images recognition of breast cancer histopathological via probabilistic neural network approach," *Journal of System and Management Sciences*, vol. 1, no. 2, pp. 53–68, 2020.

[5] L. Qin, N. Yu, and D. Zhao, "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video," *Tehnicki Vjesnik*, vol. 25, no. 2, pp. 528–535, 2018.

[6] R. H. Frank, T. Gilovich, and D. T. Regan, "The evolution of one-shot cooperation: an experiment," *Ethology and Sociobiology*, vol. 14, no. 4, pp. 247–256, 1993.

[7] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[8] S. Taasan, "One shot methods for optimal control of distributed parameter systems 1: finite dimensional control," *ICASE Report*, vol. 91, no. 2, pp. 1–20, 1991.

[9] F.-F. Li, "A Bayesian approach to unsupervised one-shot learning of object categories," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1134–1141, Paris, France, October 2003.

[10] F.-F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[11] B.-M. Lake, R. Salakhutdinov, and J.-B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6, pp. 1332–1338, 2015.

[12] B.-M. Lake, C.-Y. Lee, J. Glass, and J. Tenenbaum, "One-shot learning of generative speech concepts," *MIT Education*, vol. 36, 2014.

[13] B.-M. Lake, R. Salakhutdinov, and J.-B. Tenenbaum, "One-shot learning by inverting a compositional causal process," *Advances in Neural Information Processing Systems*, vol. 2, 2015.

[14] Z. Ji, X.-L. Chai, Y.-L. Yu, Y.-W. Pang, and Z.-F. Zhang, "Improved prototypical networks for few-shot learning," *Pattern Recognition Letters*, vol. 140, pp. 81–87, 2020.

[15] X. Liu, F. Zhou, J. Liu, and L. Jiang, "Meta-learning based prototype-relation network for few-shot classification," *Neurocomputing*, vol. 383, pp. 224–234, 2020.

[16] Z. Chen, W. Ma, N. Xu, C.-T. Ji, and Y.-L. Zhang, "SiameseCCR: a novel method for one-shot and few-shot Chinese CAPTCHA recognition using deep siamese network," *IET Image Processing*, vol. 14, no. 12, pp. 2855–2859, 2020.

[17] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," 2017, https://arxiv.org/abs/1711.04043.

[18] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia*, vol. 99, p. 1, 2020.

[19] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," 2016, https://arxiv.org/abs/1605.06065.

[20] J.-L. Lancaster, A.-R. Laird, P.-M. Fox, E.-G. David, and T.-F. Peter, "Automated analysis of meta-analysis networks," *Human Brain Mapping*, vol. 25, no. 1, p. 174, 2010.

[21] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1126–1135, JMLR. org, Sydney, Australia, August 2017.

[22] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, https://arxiv.org/abs/1803.02999.

[23] S. Chakraborty and S. Das, "k–means clustering with a new divergence-based distance metric: convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.

[24] H. A. Vrooman, C. A. Cocosco, F. van der Lijn et al., "Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification," *Neuroimage*, vol. 37, no. 1, pp. 71–81, 2007.

[25] L. Hu, J. Hu, Z. Ye, C. Shen, and Y. Peng, "Performance analysis for SVM combining with metric learning," *Neural Processing Letters*, vol. 48, no. 3, pp. 1373–1394, 2018.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep Learning Workshop*, vol. 2, 2015.
[27] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, vol. 12, pp. 3630–3638, 2016.
[28] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 3, pp. 4077–4087, 2017.
[29] M.-R. Nikoo, R.-A. Kerachian, R. A. Mohammad, and R. Mohammad, "A fuzzy KNN-based model for significant wave height prediction in large lakes," *Oceanologia*, vol. 60, no. 2, pp. 153–168, 2018.
[30] J. A. Mccarty and M. Hastak, "Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression," *Journal of Business Research*, vol. 60, no. 6, pp. 656–662, 2007.
[31] A. Jordan and M. Danijela, "Upgrading the business intelligence system by implementing the decision tree model in the *R* software package," *Studies in Informatics and Control*, vol. 29, no. 2, pp. 243–254, 2020.
[32] J. Cao, R. Panetta, S. Yue, A. Steyaert, and M.-B. Young, "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics*, vol. 19, no. 1, pp. 234–240, 2003.
[33] Y. Lin, W. Liu, and Y. Wang, "An integrated approach using cross-efficiency and shapley value in performance evaluation," *Economic Computation and Economic Cybernetics Studies And Research*, vol. 53, no. 4, pp. 209–224, 2019.

Hindawi
Complexity
Volume 2021, Article ID 6648009, 12 pages
https://doi.org/10.1155/2021/6648009

Research Article

Data Mining Algorithm for Demand Forecast Analysis on Flash Sales Platform

Mingyang Zhang,[1] **Yixin Wang**,[1] **and Zhiguo Wu**[2]

[1]*Department of Management Science and Engineering, School of Economics and Management, Beijing Forestry University, Beijing 100083, China*
[2]*Department of Logistics Management, School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Mingyang Zhang; mingyangzhang@bjfu.edu.cn

Received 26 November 2020; Revised 5 December 2020; Accepted 18 December 2020; Published 5 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

With the development of the digital economy, the emerging marketing strategy of the e-commerce flash sales has been changing the traditional purchasing habits of customers. This imposes new decision-making challenges for companies involved in flash sales. It is important for companies to build the accurate product demand forecast analysis focusing on the characteristics of the flash sales and customer behaviors. In this paper, VIPS (Weipinhui, a Chinese e-commerce platform) is taken as a case study with the key focus on how sentiment factors in customer reviews affect product demand in flash sale platforms. The paper adopts two sentiment analysis methods based on emotional dictionaries. The method with a higher evaluation index is adopted to integrate the emotional factors into the autoregressive model for product demand and assessment. The experiments prove that the autoregressive model for integrating the sentiment factors demonstrates better forecasting performances than the models without sentiment factors. The experiments further confirm that when product demand for the previous two weeks and customer review sentiment factors in the previous week are taken into consideration, demand forecast effects are most accurate.

1. Introduction

With the outbreak of COVID-19 in 2020, a new round of industrial revolution has emerged in human society. As the number of customers in e-commerce platforms keep surging, traditional customer habits have also been changing accordingly. Flash sales, originated from the French sales platform VentePrivce.com, is a special sales platform featuring the business-to-customer (B2C) e-commerce platform. The website regularly launches all sorts of famous retail products and sells them at a relatively low discounted price to the website members. Compared to normal online shopping, its strong features such as limited shopping time, quantity, and low prices are more appealing to customers [1, 2]. VIPS (VIPS is an acronym for Vipshop Information Technology Ltd.) is a pioneer in the domestic e-commerce platform sector. VIPS launched the flash sales model and is one of the most successful in operation. Since its establishment in August 2008, VIPS has been adopting the flash sales business model featuring brand discounts, flash sales, and authenticity guarantee. Many countries' e-commerce platforms have used FS platforms, e.g., HauteLook.com, Limango, and Giltand. Brands negotiate a time period with the platform for flash sales, where they sell products at a lower price and restore the original price outside of the flash sale time. For customers, such a business model mainly demonstrates the fun of panic buying as the products are sold in limited quantities at low prices. When customers secure products, they crave while others fail and they will enjoy a sense of achievement and satisfaction. So far, VIPS boasts a total of 340 million members with strong loyalty and stickiness with a repurchase rate of above 87%. Since its conception, VIPS has only spent three years in getting successfully listed on the New York Stock Exchange. Up to

now, VIPS has more than 30,000 brand suppliers in partnership. It is number 1 with 38.1% of the market share in the whole Chinese online special sales market. By the third quarter of 2019, VIPS has registered 28 consecutive profit-making quarterly performances and broken the industry record. The key to VIPS' success lies in its adoption of the flash sales model (VIPS. About us. https://www.vip.com/about-us).

Traditional flash sales websites (such as Gilt and Rue La La.) do not have sections for customer reviews. In 2019, VIPS has modified its website modules and added the customer review module. The online word-of-mouth effects greatly impact customers' decision-making, and customer reviews stand as one of the most important formats of communication [3]. Before online shoppers make their decisions, they usually resort to customer reviews and evaluations, as well as horizontal comparison to identify product information on quality so as to reduce uncertainty [4]. Therefore, reviews are regarded as one of the key drivers to future product demand [5]. Analysis of the sentimental factors in customer reviews is an important method to understand customer thinking. We take the flash sales platform of VIPS as an example to analyze how customer reviews impact flash sales' product demand.

To date, research on flash sales e-commerce is relatively limited. The literature on the flash sales e-commerce model mainly adopts the research perspective from the customer's, the retailer's, and the platform's side. The literature is predominantly focused on the model's impact on customers' decision-making and on the exploration of the psychological mechanism of the purchase urge. Also, there are very few empirical sentiment analysis studies for flash sale patterns. In addition, the research of inventory management systems mainly focuses on optimizing models and algorithms, considering factors such as logistics and location, and less on considering emotional factors [6]. Using mathematical and optimization methods, the existence of the optimal solutions is proved, and then a simple heuristic algorithm is presented to maximize total inventory profit and determine the best values of variables [7] build a systematic and improved optimization model of the supply chain inventory. They proposed ant colony algorithms and fuzzy modelling [8] develop the open-source software JSOptimizer that can be used to optimize simulation models of complex engineering systems built with JaamSim, and solve several instances of the optimization problem. Zhang et al. [9] applies a matching model of inventory control strategy for material classification in practice, and demonstrates the applicability and feasibility of the model. This paper is meant to help companies draft more accurate inventories, and restocking plans before the flash sales kickoff to prevent customer loss due to shortage of supply and to prevent inventory fee increases due to overstocking.

The research framework for this paper consists of five parts: chapter one offers an introduction to the research background and puts forward the research topic. Chapter two presents the literature review and states the foundation and bases of the research. Chapter three describes the methods and processes for reviewing data collection, states the data reviewing preprocessing method, and introduces the method of word segmentation in the end. Chapter four establishes the sentiment analysis model based on the short-term forecast. A comparison of two dictionary-based sentiment analysis algorithms are made to select the one with the better evaluation index and integrate it into the model as the sentiment factor. In the end, the short-term forecast model is established for making forecasts and conducting experiments and analyses. Chapter five summarizes the major research conclusions, makes recommendations to the platform managers and developers from the perspective of operation management, and proposes directions for future improvement for deficiencies in the paper.

2. Literature Review

Currently, there is still a relatively small number of academic researches into the flash online sales business [10]. Discovers through the empirical study that the flash sales model can further stimulate customers' desire to purchase. The less educated customers are more likely to believe in the handsome amount saved, which will drive them to place the order. A pricing strategy for factories to offer a time-bound discount which will later expire will create more benefits. Huang and Benyoucef [11] believe that the flash sales e-commerce model is beneficial for establishing brand loyalty, increasing sales, and advancing the speed of destocking. Peng et al. [12] state that the perceived value of a product is based on three criteria: the function, the emotion, and the social interaction. They discover that the perceived value has a positive correlation to purchase willingness. Time pressure works on the perspectives of perceived values emotionally and socially to create negative impacts on purchase willingness. Ferreira et al. [13] forecasts a product's future demand through machine learning and optimizes product prices on the flash sales platform. Zhang et al. [1] identified that the expectations during the sales stage depend on the reputation effect, the price for flash sales, and the inventories for flash sales.

The characteristics of a flash sales platform, such as a discount, a quantity restriction, and a time restriction are similar to the daily deals or deal-of-the-day promotions or retail outlet stores. But there are some differences between them.

(1) The first difference is the source of the product; on the flash sales platform, customers can buy products from various regions, even various countries. Groupon is a famous daily deal website. Gao and Chen [14] said "These online voucher vendors sell vouchers in specific cities at discounts ranging from 50% to 90%. These vouchers are typically offered by local businesses, such as restaurants and spas." Krasnova et al [15] mentioned that "Deal-of-the-Day (DoD) platforms have quickly become popular by offering savings on local services, products, and vacations."

(2) The deal-of-the-day or daily deal more emphasizes economies of scale; itis different from flash sales [16]. Provided that "the deal-of-the-day (or daily deal) is a group-buying website, where buyers with similar purchase interests congregate online to obtain group discounts. For interested buyers to enjoy the daily deal, the number of confirmed buyers on the particular day has to exceed the minimum required number as indicated on each website." On the FS platform, the number of confirmed buyers has no restriction.

(3) Like the daily deal website Groupon, the bricks-and-mortar shop knows how many voucher was sold; it equals the shop know a part of demand in advance. But on the FS platform, the sold quantities in the FS period are not equal to the part demand of the ricks-and-mortar shop; it will affect only the demand of the shop.

In conclusion, the literature on the flash sales e-commerce model mainly adopts the research perspective from the customer's, the retailer's, and the platform's side. The literature is predominantly focused on the model's impact on customers' decision-making and on the exploration of the psychological mechanism of the purchase urge. Despite the increasing popularity of the FS in practice, the number of literature papers on the impact of customer reviews on customer decision-making in flash sales platforms is relatively few. This paper is developed on the foundation of previous reviews. It adopts the perspective of flash sales platforms, carries out empirical studies based on real sales numbers and review data, and aims at enriching the research content of demand forecast in flash sales e-commerce models.

At the current stage, text sentiment analysis is an important research branch in the field of web data mining. It is widely used in real life. Other widely used models include: classification models [17], recommendation system [18], customer relationship management models [19], stock market prediction [20], social problems monitoring [21, 22], opinion polling [23], and competitive intelligence acquisition [24]. Sentiment analysis technologies mainly include the machine learning method and the semantic orientation method. The machine learning sentiment categorization method requires a large amount of sample training in application to set up [25]. Makes good use of the N-gram words and their special features. For the first time, they apply naïve bayes, support vector, and maximum entropy into passage-level sentiment categorization tasks. The semantic orientation method focuses on the subtraction of sentiment words and judgment of the sentiment polarity. Therefore, it does not require training beforehand [26]. Stacked denoising autoencoders (SDAs) were used to provide an infrastructure to resolve issues of sentiment recognition from textual contents. The results indicate the promising capability of SDAs to perform sentiment recognition on a multitude of domains and languages [27]. Proposes an improved stacking framework which contains multiple layers for predicting whether the stock price index will increase or decrease with respect to the price prevailing sometime earlier, if necessary, a month [28]. Build a domain-dependent sentiment dictionary, SentiDomain. They propose a weak supervised neural model that aims to learn a set of sentiment cluster embeddings from sentence global representation of the target domain. Kumar et al. [29] propose an efficient method for sentiment analysis by using particle swarm optimization, which experiments show that the proposed technique outperforms other state-of-the-art techniques. Hu and Liu [30] judge the sentiment polarity for words selected from dictionaries and complete the categorization by calculating the weighted sum.

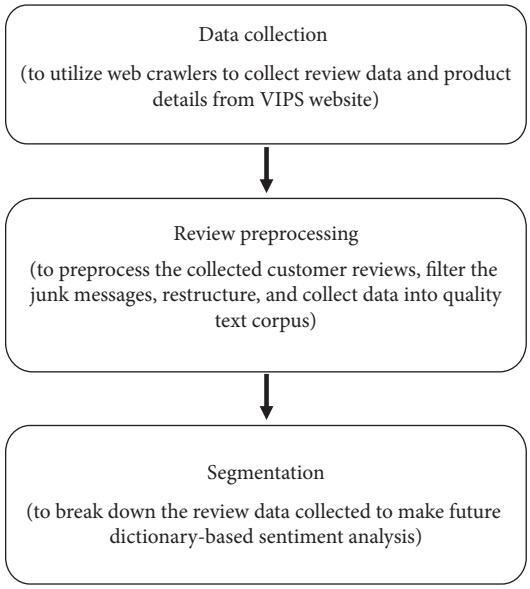
Up till now, barely any flash sales platforms have launched with the customer review block. The paper will make use of the customer review information in VIPS and apply the sentiment analysis method to produce a better demand forecast. The demand forecast model is usually realized through an autoregressive model (Water 2004), linear regression model [31], time series association analysis (Chatfield 1984), Granger causality analysis [32], and nonlinear model optimization [33]. Using the blog sentiment analysis to forecast box office sales [34], the autoregressive emotion sensitive model (ARES) was proposed. They discover that the sentiment captured in the blog one day before has the best forecasting results for the box office sales prediction for the next day [35]. It is discovered that when the film is released, heated discussion will take place in the microblog platform. Later, the number of relevant blogs will gradually decrease. The box office returns go through similar processes. Return on the Sunday of the release week is usually the highest. Therefore, they construct a forecast model based on linear regression to forecast box office return for the week of release [36]. A linear regression empirical study was carried out for audio, video plays as well as electronic cameras in Amazon [31]. Forecast transmission of infectious diseases by setting up a logistic regression model. Using blog mention count to forecast peak sales for books [37], we use a series of predictive classifiers such as Light GBM, XGBoost, Logistic Regression, and Random Forest in order to evaluate the probability of a customer entering loan default. Gruh et al. [38] utilize time series correlation analysis to forecast the timing advance. They found that different books in the samples have different timing advances. This may have to do with the fact that the arrival of a book's peak sales is subject to the occurrence of various social events. Forecasting stock market development using blog sentiment analysis [39], the Granger causality analysis was applied to analyze various sentiment time series and the Dow Jones index time series. They believe that the best timing advance for the forecast is two days [40]. The impacts of product reviews in a competitive market were proved [41]. It is identified that the sentiment value in product reviews has a significant impact over future product demand [42]. The convention rate of the influence was studied from sentiment value in online reviews.

The above-mentioned research shows that the autoregressive model, as a practical model for addressing the problem of time series, has already been widely applied to various forecast scenarios. A short-term sentiment aware

autoregressive model (SAAR) can be established based on the sentiment factors and previous sales. The paper mainly investigates sentiment analysis methods to study social behavior, emotional dictionaries for flash sales patterns. Authentic product sales numbers and review data are adopted to verify the model and guarantee authenticity and accuracy of the research.

3. Data Collection and Data Processing

There are three steps for data collection and data processing:



The paper uses GooSeeker as the tool for data collection and selects the domestic appliances categories in the VIPS platform for data mining. In every category, 2 brands from 17 types of products will be reviewed (in order to clearly figure out what impact of consumer reviews on product's demand, we only consider the two brander case). In regards to reviewing data, the collection is done in a reverse manner. In other words, data are collected from the very day on and backward. For each one of the products, data mining will end when the date reaches December 30th. In total, 10,000 reviews have been collected. In terms of the product details, the paper has collected the real-time domestic appliances popularity ranking list from December 30th, 2019, to March 29th, 2020. Product details are collected on a regular daily basis, which has amounted to 1547 items. The data mining rules in GooSeeker's MS station have been adopted and customized into those of VIPS' own. For each review, the time, customer name, customer level, and review text have all been noted.

As many low-quality items in text reviews may affect future analyses, five steps need to be taken in the preprocessing stage as follows:

(1) Removing duplication: removing duplication is to delete repetitive messages in the customer review. The same person may purchase multiple times in one store, which leads to repetitive reviews. In such circumstances, only the earliest review will be saved and the remaining repetitive ones will be deleted.

(2) Mechanical compression: this step is to process repetitive parts in the sentence. In the paper, the redundant part of the text is processed, mainly centering on the beginning and the end of sentences. For instance, in "Thumbs up. Not baaaaaaaaaaaaaaaaad." only "bad" needs to be reserved. Otherwise, the future sentiment value calculation will not be affected. As a result, the whole sentence is compressed into "Thumbs up. Not bad."

(3) Short sentence compression: short sentence compression is mainly about deleting extremely short or meaningless reviews. In this paper, texts with less than five international characters are deleted. Short sentences include sentences that are short to start with and those that become short after mechanical compression, i.e., those long texts featuring meaningless repetitiveness.

(4) Removing emojis and emoticons: the last step is mainly about manually deleting emojis and emoticons in sentences.

(5) For Chinese sentences: segmentation will be imposed on the word series. After segmentation, the sentence 'I am very satisfied. Looking Good. I'm so into it.' will be turned into '/I'm/very/satisfied/, Looking/Good. I'm/so/very/into/it.' After removing stop-words, the review will look like 'Very/satisfied. Not bad/So/into it.'

In order to judge whether there are sentiment words contained, we need to segment every review and accurately keep the keywords. Accuracy of the segmentation means a lot to the following analysis. Therefore, the method with better effects needs to be selected. For Chinese sentence segmentation, many methods are available. This paper has resorted to '*Jieba*' and Python's Chinese segmentation package and to handle reviews in text documents. Accuracy, efficiency, generality, and applicability are the most important factors in segmentation performances. This system offers more than 97% accuracy [43] and features easy installation, extensive language support, and quite a degree of popularity. After segmentation, unnecessary words need to be removed. Unnecessary words include prepositions, pronouns, function words, and characters irrelevant to sentiment analysis. After preprocessing, reviews for the 17 products will be segmented.

Thanks to '*Jieba*' above 97% accuracy [43], the approach is used in this paper to segment the texts. The paper adopts the generative model based on the *Snownlp* stop-word list (https://github.com/isnowfy/snownlp). The negative words and degree level adverbs are filtered to generate a new stop-word list.

4. Sentiment Aware Model (SAM) for Short-Term Forecast

4.1. Model Assumption. First of all, the relationship between sentiment value and product demand in product reviews is investigated. The experimental data are used to test such a correlation. Then, the model assumption is proposed.

The paper takes the domestic appliances in the VIPS flash sales platform as the target of research and collects customer reviews during a designated period. However, as there is no direct access to the sales figures, the number of reviews published can be roughly taken as the demand for the domestic appliances.

Here it is assumed as follows:

H1: the number of reviews for a product equals the demand for the corresponding product.

Due to the unique nature of the flash sale platform, we cannot see the sales quantities of the product on VIPS, and the platform does not show bad reviews, only positive ratings and the number of product reviews. Moreover, Park et al. [44].found that purchasing intention increased as the number of reviews increases. Therefore, we made the assumption of H1.

Product review mining is an important application of sentiment analysis. Scholars adopt various econometric models and research methods to measure the enterprise communication effects of online product reputation in multiple dimensions. The three most commonly used dimensions are volume, valence, and dispersion. Volume is mainly referring to the number of customer reviews for a certain product. It reflects the awareness effect of the online reputation [45]. According to the rules of VIPS, reviews will only be shown when the total number reaches 999. Therefore, there is no direct access to the total review number. That's also why the impact of volume on sales is not considered for now. Dispersion means the degree of communication in different online communities. The higher the dispersion is, the greater the influence. Because this paper only looks at the VIPS community, measuring dispersion is not applicable. Therefore, valence is the major dimension used to analyze the impact of online reputation on product demand.

Valence measures the customers' feedback on the products in both good and bad, positive and negative ways and is usually measured with an overall score (good-bad) or a ratio between the good and bad (good/bad). It reflects the persuasive effect of a product's online reputation [46]. It is discovered that improvement of book reviews can increase demand for the book. In the meanwhile, drops in demand caused by negative reviews are more prominent compared to the increase in demand incurred by positive reviews. Similarly, Floh et al. [47] found that the stronger a review is, the more likely it will stimulate purchase increases or decreases. In other words, intense positive or negative reviews create greater influence than those with mixed emotions.

Based on the analysis above, the following assumption is made:

Table 1: Sale level and comments emotion score.

	Sale level	c
Week 1	70	8.58
Week 2	57	9.21
Week 3	63	9.77
Week 4	34	6.59
Week 5	38	6.67
Week 6	69	6.59
Week 7	67	11.40
Week 8	117	17.66
Week 9	109	18.15
Week 10	112	10.59
Week 11	93	9.36
Week 12	78	10.22
Week 13	87	7.35

H2: a reviewer's overall emotion (demonstrated through written feedback) toward the product creates a positive or negative impact on product demand.

The demand data have been acquired for "Media 304 Stainless Domestic Electric Kettle 1512d" in 13 weeks, in 944 review messages, through the web crawler technology. Using the improved SAM, the sentiment value in product reviews is calculated as shown in Table 1.

The SAAR model proposed by Liu et al. [34] reveals that the sentiment information captured in blogs can achieve the best performance in forecasting the film box office ticket sales for the next day. Using the number of mentions in blogs to predict books' peak sales volumes [38], the time series correlation analysis was utilized to confirm the advancement time for the prediction. They identify differences between books. However, basically, the gap ranges from several days to several weeks. Therefore, the time lag is uncertain, and the advancement is usually affected by the scope of application, social behavior, and the method of sentiment analysis. The current sentiment value may have an impact on the product demand for the next cycle. When the correlation analysis is carried out, the impact from the sentiment value is delayed. If it is assumed that the time lag is one cycle [34], then that means sentiment value in the first week affects demand in the second week, and so on so forth. Therefore, the data to be analyzed for correlation is $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, and the formula for calculating the ratio r is shown in the following formula:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}}. \quad (1)$$

In the formula, x_i is the demand for week i and $\overline{x}$ is the average demand within the observation period. y_i is the sentiment value in week i and $\overline{y}$ is the average sentiment value within the observation period.

The calculation results are shown in Table 2.

Such experiment results to demonstrate a strong correlation between the sentiment value and demand at the 0.01 level (two sides).

Based on the analysis above, the following assumption is made:

Table 2: Correlation analysis results.

		Demand	Sentiment value
Demand	Pearson correlation	1	0.721**
	Significance (double tail)		0.008
	Number of cases	12	12
Sentiment value	Pearson correlation	0.721**	1
	Significance (double tail)	0.008	
	Number of cases	12	12

Table 3: Example of sentence sentiment score.

Vocabulary	Sentiment score
Can't turn around (hui tian Wu li)	−4.74492768973
Speechless (wuyu)	−3.66149641358
Anxious (ganzhao ji)	−2.17885364262
No second to (buyayu)	0.858219035923
Bravo (piao liang)	2.23552351188
Super value (chao zhi)	3.44664867894

Table 4: Degree level word weights.

Degree level words	Example	Weight
"Most"	One hundred percent and extremely	2
"Pretty"	Pretty and very much	1.5
"Quite"	Quite	1.25
"Rather"	Rather	0.5
"A little"	Not quite and a little	0.25
"Too"	Too and over	0.1

H3: current sentiment value impacts demand during the next cycle.

4.2. Dictionary-Based Sentiment Analysis Method. Dictionaries for emotions and degree level words have been compiled. First of all, the *Boson NLP* emotional dictionary is selected to judge the sentiment scores. Negative scores represent more negative words. Positive scores represent more positive words. The degree of emotions can be reflected through the scores. Table 3 shows examples of words and their sentiment scores based on the *Boson NLP* emotional dictionary.

This paper uses a degree level adverb dictionary and the integrated negative word dictionary from the sentiment analysis vocabulary (beta version) in the cnki network. Customers often resort to degree level adverbs and negative adverbs in expressing emotions. For instance, they may use degree level adverbs (e.g., 'quite', 'extremely', 'somewhat', 'a little') to emphasize the subtle differences in emotions. Also, some negative adverbs like "not" will change the sentiment polarity. The sentence "she is not beautiful." is an example. The degree-level word list is shown in Table 4. Based on the relevant word information provided by cnki, a certain weight is given to the common degree level adverbs in the corpus. According to cnki, the degrees are categorized into six levels: extremely (most), pretty, quite, rather, a little, and too. The weight given in this paper is noted as follows:

W is set as the weight of the degree level word and S as the sentiment word value. Sentiment_index is the subscript for the sentiment word. Calculation of the sentiment degree is as follows:

According to the above Algorithm 1, for the example of "I'm satisfied. The look is pretty good." The calculation of the sentiment in the sentence works as follows:

(1) The words 'satisfied', 'pretty', and 'good' are left after preprocessing the data and removing stop-words like 'I'm', 'look', and 'is.'

(2) The sentiment words 'satisfied' and 'good' have weights of 2.84 and 2.65. The degree level word 'pretty' has a degree value at 1.52.

(3) Because there is no negative word, the negative sentiment value is 0.

(4) Therefore, the sentiment value of the sentence is $2.84 + 2.65 * 1.52 = 6.868$.

In the end, the experiment effectiveness is evaluated. Because the sentiment analysis software for *ROSTCM6* is based on optimizing emotional dictionaries, the accuracy is higher than those based on word vectors or neural networks. This paper uses *ROSTCM6* in analyzing the sentiment value in customer reviews. For *ROSTCM6* software, the experiment results include positive emotion, negative emotion, and neutral emotion. In this paper, if the sentiment value of the sentence is greater than 0, then it means positive emotion. If the sentiment value is less than 0, then it means negative emotion. The benchmark is marked out manually based on the condition that it is all correct and does not involve individual differences. The closer the result is to the benchmark, the more accurate the model proves to be.

Three major assessment indexes are adopted here: recall, precision, and F-measure.

(1) *Recall rate*: investigating the comprehensiveness of the sentiment categorization model and reflecting the ratio of the number of correctly identified to the number of identified total after the experiment.

$$\text{Recall rate} = \frac{\text{correctly identified ones}}{\text{identified total}}. \quad (2)$$

(2) *Precision rate*: investigating accuracy of the model and reflecting the ratio between the number of the correctly identified against the number that ought to be identified after the experiment.

$$\text{Precision rate} = \frac{\text{correctly identified ones}}{\text{actual total}}. \quad (3)$$

(3) *F-measure*: the harmonic mean of the two when the recall rate and the precision rate are viewed as equals.

$$F - \text{measure} = \frac{(2 * \text{recall rate} * \text{precision rate})}{(\text{recall rate} + \text{precision rate})}. \quad (4)$$

The paper experiments with 100 reviews for the first product. From results in Table 5, it can be noted that the tree

```
(1) for screening the segmentation result, do
(2)     if the word belongs to the sentiment vocabulary, then
(3) score+ = W ∗ S.
(4) sentiment_index+ = 1
(5)     if sentiment_index is smaller than the total amount of all sentiment word, then
(6) for degree level adverbs or negative words that exist between the current sentiment word and the next, do
(7) if it is a negative word, then
(8) W* = −1.
(9) end if
(10) if there is any degree level word, then
(11) W* = V
(12) end if
(13) end for
(14) end if
(15) end if
(16) end for
```

Algorithm 1

assessment indexes in the new model are all higher than those in the original model. The optimized model shows significant performance improvement. The paper adopts the current sentiment analysis model to develop SAM based on short-term forecasts.

4.3. Model Development. Sales data are used instead of demand data to predict current product demand using previous sales performances. In real life, the current sales of a product show a certain correlation with its previous sales. Therefore, the autoregressive model is more suitable. The domestic appliances in VIPS are our research object. Customer reviews for the domestic appliances in VIPS are collected for a certain period. As there is no direct access to the product sales numbers, the number of reviews for a product is taken as the approximate number of product sales. Affected by multiple factors such as the seller preparation time, the delivery speed, postponement of buyers' feedback, and looking at the data on a daily basis, there may be days when there are zero reviews or a huge amount of reviews. Data with such big fluctuations are apparently not applicable for model development. In order to reduce impacts from fluctuations, we take weeks as the time series unit. The demand for the product over 13 weeks in VIPS is captured as follows:

According to the autoregressive distributed lag, it is forecasted that the demand for the domestic appliance during the time-frame X_t requires p periods before X_t remains stable. Otherwise, different treatment will be carried out. As in Figure 1, the demand features prominent fluctuations during different phases and different treatment is necessary. By calculating the logarithm for each element in the demand series $\{x_t\}$, a new demand series $\{y_t\}$ is produced. Please refer the following formula:

$$y_t = \log_2(x_t). \tag{5}$$

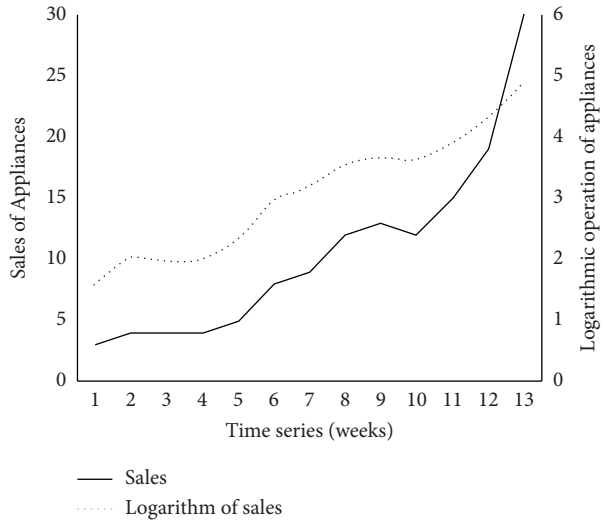


Figure 1: Example of domestic appliance sales

After processing the data, the autoregressive coefficient must be forecasted, and then, the observed data must be fit into a linear parameter model. In this paper, the estimate will be carried out upon the training group by ordinary least square. In the end, the model is used to analyze and investigate relations between demand data in different phases to forecast current demand.

To apply the autoregressive distributed lag (ADL) model, ADF and other relevant tests need to be passed. To secure the stability of the data, an ADF test needs to be applied to the new demand series. In this paper, the critical value of 0.05 is taken as the standard. If the data fails the test, they will be directly abandoned. After the ADF test, 12 categories of domestic appliances pass. The autocorrelation (AC) coefficient is to be calculated for the new demand series $\{y_t\}$ after the ADF test. AC coefficient between variables is needed for

the model. If the coefficient is greater than 0.5, the AC model should be set up. Otherwise, the data should be abandoned. After the test of the AR coefficient, 10 domestic appliances survive.

Besides previous sales performance of the product, customer opinions also have influence on the current sales. Therefore, the sentiment factor is brought in to optimize the model. If we take C_t as the number of product reviews during the observation period t and set the observation cycle at one week, the average sentiment value during t period is defined as in the following formula:

$$S_t = \frac{\sum_{c=C_t} e}{|C_t|}, \tag{6}$$

S_t is the average sentiment value during the observation period t. e is the value after *ROSTCM6* is used to generate sentiment analysis. S_t will be integrated with the autoregressive model to obtain a short-term-based SAM as shown in formula (7). In essence, it is an application of the ADL model.

$$y_t = \sum_{i=1}^{p} \theta_i y_{t-i} + \sum_{j=1}^{q} \lambda_j S_{t-j} + \varepsilon_t. \tag{7}$$

y_t is the product sales as a function of time t. S_t is a sentiment element function of t. q and p are parameters selected by the users. Parameter q is selected by users, and it is sentiment information from a few weeks ago, while p is sales information from a few weeks ago. θ_i is the demand coefficient in history, λ_j is the sentiment coefficient, and ε_t is the error term (white noise with an average value of 0).

4.4. The Model Experiment Results and Analysis. The paper carries out the ADF test and autocorrelative examination after data processing. It selects demand data that can be used, which is data for ten domestic appliances in VIPS over 13 weeks and categorizes them into the training group and test group. In the training group, the study is carried out toward the coefficient θ_i $(i = 1, 2, \ldots, p)$ and λ_i $(i = 1, 2, \ldots, q)$ in the model by ordinary least squares.

The paper evaluates the model effectiveness with mean absolute percentage error, MAPE. The calculation of MAPE is shown in the following formula:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{n} \frac{\left|\text{tred}_i - \text{true}_i\right|}{\text{true}_i}. \tag{8}$$

In the formula i is week number, $(i = 12, 13, \ldots, n)$, n is the estimated time in total, Pred_i is the estimated value obtained using the model. True_i is the actual value. The smallerthe MAPE is, the better the model is in making forecasts.

The paper conducts the ADF test using *EVIEWS10*. Only those that withstand the test can be applied to the autoregressive model with the parameters filled in place to generate the forecast value. After processing, the MAPE value can be generated. The experiment is carried out on the condition of $p \in [1, 6]$ and $q \in [1, 6]$. The combination of p and q with the best result is selected. The data for the first 11 weeks are taken as the training group, the data for the last two weeks are taken as the test group, and then a forecast is made for every domestic appliance in the two weeks test period. First, parameter q in the optimized model is fixed at 1 and tests are carried out on the condition that $p \in [1, 6]$. Results are shown in Figure 2.

As is shown in Table 6, when p equals 6 or 5, the Sam outperforms the autoregressive model. When p equals 1, 2, 3, and 4, the forecast effect is even more prominent. It shows certain relations between the forecast effect and the sentiment factors.

When p equals 5, the effect of the SAM is not easily observed. When p equals 6, the autoregressive model has better performances than the SAM. When p equals 5 and 6, the average value of MAPE in the SAM is bigger than that in the autoregressive model. This might be a result of influence from sentiment words in the reviews. In the autoregressive model, the MAPE values for products at 2, 7, and 8 are about 7% lower than those in the independent AR model. Upon observation of the data, neutral emotions take up 1/3 of the total, accounting for quite a prominent percentage. The major reason is that within a certain cycle t, and the neutral emotions contained in sentiment values in product reviews only increases the number of reviews. It is for sure that the ultimate sentiment information will be weakened. However, in real life, the neutral reviews often contain relatively complicated messages, details of which cannot be easily processed, and generate biased results.

When p equals 2, effect of parameter q on the model is shown in Figure 3. The result shows the worst performance of the model when q equals 1. That means that the sentiment information in reviews published one week before has the best effect for forecasting demand.

Figure 4 shows the overall situation for ten different electronic products. It can be noted from the figure that, in the autoregressive model and the SAM, when $p = 2$ and $p = 6$, the best and worst cases, respectively, happen. It means the selection of p itself also affects accuracy of the model. If the value of p is too small, the hidden relationship between numbers can be easily ignored. However, when the value of p is too big, there will be too much distracting data. In the experiment, when $p = 2$, the best effect is achieved, which means sales of a certain week is affected by sales of the previous two weeks. It has to do with the time and frequency of flash sales. The research object of the paper has been arranged for flash sales promotion in weeks 6 and 7 and weeks 10 and 11. Sales performances in the week before flash sales and in the first week during flash sales both have impacts on sales in the second week during the flash sales period. That means the reputation of the product itself and

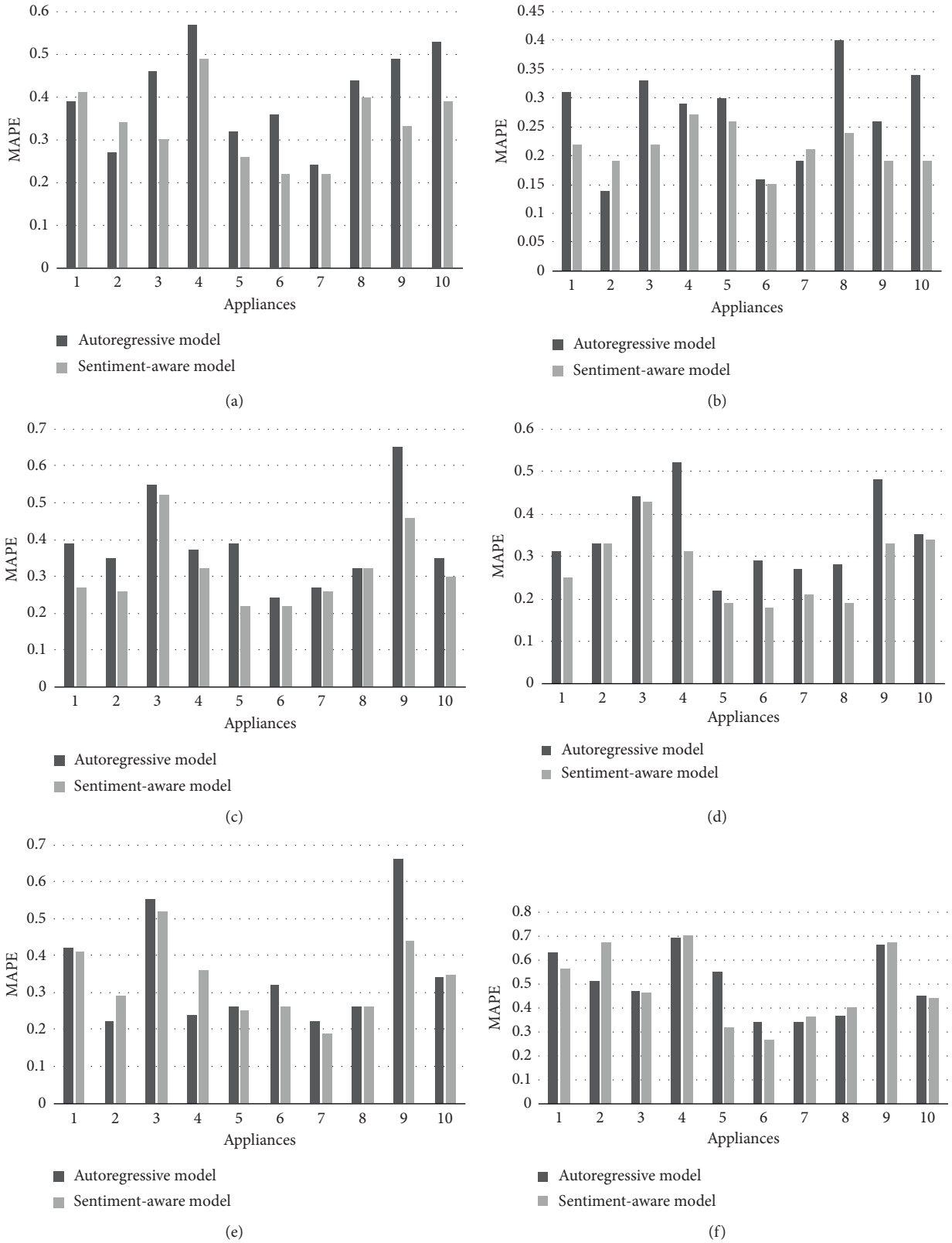


Figure 2: Experimental results of appliances sales forecast, (a) results of sales information one week ago, (b) results of sales information two weeks ago, (c) results of sales information three weeks ago, (d) results of sales information four weeks ago, (e) results of sales information five weeks ago, and (f) results of sales information six weeks ago.

the format of flash sales are both influencing factors. For products with a good reputation, when the flash sales season comes to the end, customers will feel the time pressure and get further stimulated to purchase. For products with a bad reputation, a joint effect can still be noted when the flash sales period is about to end. Furthermore, the sales performance in the first and second week during the flash sales period also affect sales performance during regular time.

The experiment result shows that sentiment information can affect demand. When the sentiment factors are considered in the autoregressive model, the forecast effects register a prominent improvement. In inventory management, the forecast can be utilized to predict the number of orders for the next cycle to control an enterprise's overall stock and purchase numbers and to reduce the inventory cost. Regarding sales management, the forecast model should be first used for prediction. If huge fluctuations are identified, it is highly likely that certain sales strategies have been applied to competitive products. The use of the model is beneficial for the company as it can adjust the product price based on the competitive prices in a timely manner.

5. Discussion and Conclusions

The paper takes the VIPS as the object of research and prioritizes investigating how sentiment factors in customer reviews affect demand forecasts for products in the flash sales platform. The contribution of the paper is mainly summarized in the following three aspects:

(1) Based on the authentic sales data and review data from the flash sales platform, explorations are made regarding influencing factors on customer behavior in flash sales platforms. Correlation between sentiment factors and demand is proved through solid experimental results.

(2) A science-based analysis framework is offered to enterprises when they establish sentiment-oriented analysis model for product sales. The paper has adopted theories and methods related to a sentiment analysis and implemented sentiment mining on customer review data. Considering the special words and factors that may affect the customers' reviews sentiment analysis results, text data have been converted to numerical data to improve the original sentiment analysis model and increase its accuracy.

(3) Investigations are carried out to use the flash sales model in the correct way to forecast demand and enhance enterprise performances, especially for inventory optimization. The experiment has proved that the autoregressive model, which integrates the sentiment factors' features, leads to better forecast. Furthermore, the autoregressive model has best performances in terms of demand forecast, driven by customers' sentiment factors, when the forecast is targeted at one or two weeks beforehand.

Table 5: Comparison of evaluation indexes.

	Original model	Optimized model
Recall (positive)	0.87	0.95
Precision (positive)	0.76	0.91
Precision (negative)	0.31	0.53
Precision (negative)	0.50	0.69
Recall	0.59	0.74
F (positive)	0.81	0.93
F (negative)	0.38	0.60

Table 6: Analysis of experiment results.

p	Percentage of SAM outperforming auto regressive model (%)	Average MAPE value in SAM against that in auto regressive model
1	80	0.85
2	80	0.85
3	90	0.83
4	90	0.80
5	60	1.01
6	50	0.97

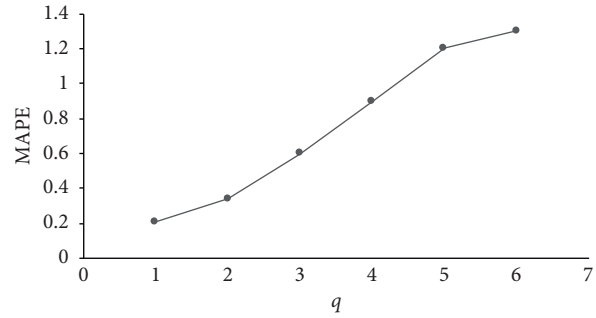


Figure 3: Effect of parameter q on model.

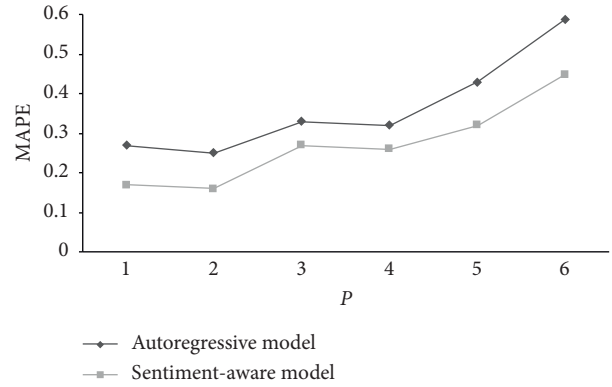


Figure 4: Result of comparative experiment.

As for future research, the following is proposed:

(1) The Internet catchphrases and emojis have already been quite pervasive in nowadays reviews, which can also demonstrate the emotional attitude of the customers. The paper does not have dedicated research into emojis and only filters the relevant

information. In later stages, sentiment analysis can be carried out for both the trendy phrases and emojis.

(2) The quality of reviews is not taken into consideration, and the false reviews cannot be fully filtered. Therefore, the authenticity of the reviews can further be improved. In later stages, customer levels and thumbs-up numbers for reviews can also be considered to optimize the model.

(3) This paper adopts the time series data. However, the future study can resort to panel data and select multiple factors to establish the model for the empirical study of product demand.

Data Availability

The data used to support the findings of this study are currently under embargo, while the research findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work. The authors declare that they do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant no. 71901027.

References

[1] M. Zhang, T. C. E. Cheng, and J. Du, "Advance selling of new products to strategic consumers on flash sale platforms," *International Journal of Logistics Research and Applications*, vol. 21, no. 3, pp. 318–331, 2018.

[2] M. Zhang, J. Zhang, T. C. E. Cheng, and G. Hua, "Why and how do branders sell new products on flash sale platforms?" *European Journal of Operational Research*, vol. 270, no. 1, pp. 337–351, 2018.

[3] B. Bickart and R. M. Schindler, "Internet forums as influential sources of consumer information," *Journal of Interactive Marketing*, vol. 15, no. 3, pp. 31–40, 2001.

[4] F. Zhu and X. Zhang, "Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics," *Journal of Marketing*, vol. 74, no. 2, pp. 133–148, 2010.

[5] N. Hu, L. Liu, and J. J. Zhang, "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Information Technology and Management*, vol. 9, no. 3, pp. 201–214, 2008.

[6] R. Keshavarzfard, A. Makui, and R. Tavakkoli-Moghaddam, "A multi-product pricing and inventory model with production rate proportional to power demand rate," *Advances in Production Engineering & Management*, vol. 14, no. 1, pp. 112–124, 2019.

[7] W. Yu, G. Hou, and J. Li, "Supply chain joint inventory management and cost optimization based on ant colony algorithm and fuzzy model," *Tehničkivjesnik*, vol. 26, no. 6, pp. 1729–1737, 2019.

[8] D. Katsios, A. S. Xanthopoulos, D. E. Koulouriotis, and A. Kiatipis, "A simulation optimisation tool and its production/inventory control application," *International Journal of Simulation Modelling*, vol. 17, no. 2, pp. 257–270, 2018.

[9] Z. L. Zhang, Y. F. Wang, and Y. Li, "Inventory control model based on multi-attribute material classification: an integrated grey-rough set and probabilistic neural network approach," *Advances in Production Engineering and Management*, vol. 14, no. 1, pp. 1854–6250, 2019.

[10] D. Mattioli, "Macy's plan: boots, bieber," *The Wall Street Journal*, pp. 26-27, 2011.

[11] Z. Huang and M. Benyoucef, "User preferences of social features on social commerce websites: an empirical study," *Technological Forecasting and Social Change*, vol. 95, pp. 57–72, 2015.

[12] L. Peng, W. Zhang, X. Wang, and S. Liang, "Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: considering the impact of product involvement," *Information & Management*, vol. 56, no. 2, pp. 317–328, 2019.

[13] K. J. Ferreira, B. H. A. Lee, and D. Simchi-Levi, "Analytics for an online retailer: demand forecasting and price optimization," *Manufacturing & Service Operations Management*, vol. 18, no. 1, pp. 69–88, 2016.

[14] F. Gao and J. Chen, "The role of discount vouchers in market with customer valuation uncertainty," *Production and Operations Management*, vol. 24, no. 4, pp. 665–679, 2015.

[15] H. Krasnova, N. F. Veltri, K. Spengler, and O. Günther, ""Deal of the day" platforms: what drives consumer loyalty?" *Business & Information Systems Engineering*, vol. 5, no. 3, pp. 165–177, 2013.

[16] Y. Liu and J. Sutanto, "Buyers' purchasing time and herd behavior on deal-of-the-day group-buying websites," *Electronic Markets*, vol. 22, no. 2, pp. 83–93, 2012.

[17] O. Voican, "Using data mining methods to solve classification problems in financial-banking institutions," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 54, no. 1, 2020.

[18] J. Yoon and S. Joung, "A big data based cosmetic recommendation algorithm," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 40–52, 2020.

[19] D. Kurunathan, S. Shanmugathas, and K. Ashoka, "Analysis of relation between customer behavior and information Technology market," *Journal of System and Management Sciences*, vol. 9, no. 1, pp. 87–104, 2019.

[20] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications*, vol. 32, pp. 9713–9729, 2019.

[21] J. Awwalu, A. A. Bakar, and M. R. Yaakub, "Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter," *Neural Computing and Applications*, vol. 31, no. 12, pp. 9207–9220, 2019.

[22] L. Fu and Y. Dong, "Research on internet search data in China's social problems under the background of big data," *Journal of Logistics, Informatics and Service Science*, vol. 5, no. 2, pp. 55–67, 2018.

[23] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, no. 65, pp. 3–14, 2017.

[24] J. Roca-Gonzalez, J. Vera-Lopez, and G. Rodriguez-Bermudez, "Analysis of patent# US2014/0319274a1: a case study of simulations for new designs review," *International Journal of Simulation Modelling*, vol. 17, no. 3, pp. 405–418, 2018.

[25] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint cs/0205070, 2002.

[26] H. Sagha, N. Cummins, and B. Schuller, "Stacked denoising autoencoders for sentiment analysis: a review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 5, Article ID e1212, 2017.

[27] M.-Q. Jiang, J.-P. Liu, and L. Zhang, "An improved stacking framework for predicting stock price index direction," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, pp. 187–202, 2019.

[28] M. Ahmed, Q. Chen, and Z. Li, "Constructing domain-dependent sentiment dictionary for sentiment analysis," *Neural Computing and Applications*, vol. 32, pp. 14719–14732, 2020.

[29] R. Kumar, H. S. Pannu, and A. K. Malhi, "Aspect-based sentiment analysis using deep networks and stochastic optimization," *Neural Computing and Application*, vol. 32, pp. 3221–3235, 2020.

[30] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, Washington, DC, USA, August 2004.

[31] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[32] J. R. Freeman, "Granger causality and the times series analysis of political relationships," *American Journal of Political Science*, vol. 27, no. 2, pp. 327–358, 1983.

[33] Y. Yang and Z. Duan, "An effective co-evolutionary algorithm based on artificial bee colony and differential evolution for time series predicting optimization," *Complex and Intelligent Systems*, vol. 6, no. 4, 2020.

[34] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: a sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 607–614, Amsterdam, The Netherlands, July 2007.

[35] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492–499, Toronto, Canada, August 2010.

[36] N. Archak, A. Ghose, and P. G. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management Science*, vol. 57, no. 8, pp. 1485–1509, 2011.

[37] A. Coşer, M. M. Maer-Matei, and C. Albu, "Predictive models for loan default risk assessment," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, 2019.

[38] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 78–87, San Diego, CA, USA, August 2005.

[39] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[40] W. Jabr and Z. Zheng, "Know yourself and know your enemy: an analysis of firm recommendations and consumer reviews in a competitive environment," *MIS Quarterly*, vol. 38, no. 3, pp. 635–654, 2014.

[41] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: a case study in the movie domain," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 4, pp. 720–734, 2010.

[42] S. Ludwig, K. De Ruyter, M. Friedman, E. C. Brüggen, M. Wetzels, and G. Pfann, "More than words: the influence of affective content and linguistic style matches in online reviews on conversion rates," *Journal of Marketing*, vol. 77, no. 1, pp. 87–103, 2013.

[43] L. Zhang, *Python Data Analysis and Mining Practice*, Mechanical Industry Press, Beijing, China, 2016.

[44] D.-H. Park, J. Lee, and I. Han, "The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement," *International Journal of Electronic Commerce*, vol. 11, no. 4, pp. 125–148, 2007.

[45] S. Moon, Y. Park, and Y. S. Kim, "The impact of text product reviews on sales," *European Journal of Marketing*, vol. 48, pp. 2176–2197, 2014.

[46] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006.

[47] A. Floh, M. Koller, and A. Zauner, "Taking a deeper look at online reviews: the asymmetric effect of valence intensity on shopping behaviour," *Journal of Marketing Management*, vol. 29, no. 5-6, pp. 646–670, 2013.

Hindawi
Complexity
Volume 2020, Article ID 6612043, 12 pages
https://doi.org/10.1155/2020/6612043

Research Article

Network Pseudohealth Information Recognition Model: An Integrated Architecture of Latent Dirichlet Allocation and Data Block Update

Jie Zhang, Pingping Sun, Feng Zhao, Qianru Guo, and Yue Zou

School of Economics and Management, Shandong University of Science and Technology, Qingdao 266590, China

Correspondence should be addressed to Feng Zhao; chinazhaof@163.com

Received 22 November 2020; Revised 5 December 2020; Accepted 10 December 2020; Published 21 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

The wanton dissemination of network pseudohealth information has brought great harm to people's health, life, and property. It is important to detect and identify network pseudohealth information. Based on this, this paper defines the concepts of pseudohealth information, data block, and data block integration, designs an architecture that combines the latent Dirichlet allocation (LDA) algorithm and data block update integration, and proposes the combination algorithm model. In addition, crawler technology is used to crawl the pseudohealth information transmitted on the Sina Weibo platform during the "epidemic situation" from February to March 2020 for the simulation test on the experimental case dataset. The research results show that (1) the LDA model can deeply mine the semantic information of network pseudohealth information, obtain the features of document-topic distribution, and classify and train topic features as input variables; (2) the dataset partitioning method can effectively block data according to the text attributes and class labels of network pseudohealth information and can accurately classify and integrate the block data through the data block reintegration method; and (3) considering that the combination model has certain limitations on the detection of network pseudohealth information, the support vector machine (SVM) model can extract the granularity content of data blocks in pseudohealth information in real time, thus greatly improving the recognition performance of the combination model.

1. Introduction

At present, pneumonia caused by the new coronavirus has been effectively controlled nationwide, but the panic and fear caused by it are making people nervous. People are attempting to find various effective methods for improving their immunity to resist virus invasion and prevent virus infection by the new coronavirus. Under this background, some people take advantage of the panic mentality of the public to produce and disseminate a large amount of pseudohealth information on the Internet in the name of health. For instance, "drinking strong liquors can kill novel coronavirus," "drinking radix isatidis and smoking vinegar can prevent novel coronavirus," "drinking sterilizing fluid can kill novel coronavirus," and "wearing multilayer masks can prevent novel coronavirus." The publishers and disseminators of this pseudohealth information, with personal interests, act as "unhealthy" in the name of "healthy" and induce unwise behaviors of people who do not know the truth, which has brought great harm to the physical and mental health of the general public; additionally, it will cause property loss and life danger. All kinds of "Health Articles," "Cancer Alert," and "Private Sector" are filled in WeChat's circle of friends. Not only on social platforms, but also the whole network environment highlights a serious problem: health information is full of all kinds of health care pseudoscience, and the information is enough to make some people who lack health knowledge and literacy believe this kind of pseudohealth information. In addition, pseudohealth information spreads unscrupulously in rural areas, resulting in a series of serious consequences. For example, in recent years, there have been activities to promote fake health care products in rural areas

in China. The swindlers take advantage of the mentality of rural residents, such as seeking cheap prices and worrying about health, to carry out swindling actions which results in heavy losses to farmers. In 2015, Financial Channel of China Central Television reported that acetochlor pesticide residues were detected in strawberries, and long-term consumption would cause cancer risks. For this kind of pseudohealth information, it is difficult for nonprofessionals to distinguish whether the information is true or false. Although professionals interpreted that dosage determines toxicity with eight validation samples, it still caused a large scale of unsalable strawberries and brought great economic impact to farmers. Therefore, effective identification of pseudohealth information in networks is of great significance for maintaining the physical and mental health of the general public.

At present, there is no universally accepted definition of "pseudohealth information" in society. In general, pseudohealth information is interpreted as false health information without a factual basis, but in the real world, much pseudohealth information is fabricated based on certain facts, which only extends, distorts, exaggerates, and even fabricates the facts. Therefore, the pseudohealth information to be studied in this paper is that fabricated without a factual basis or with a certain factual basis but distorted or exaggerated by the publisher, the so-called health information that deviates from the truth. Network pseudohealth information refers to false health information that is fabricated or distorts the truth transmitted specifically through social media on the network. It is the "noise" in health transmission; it often induces people to form incorrect health cognition and even engage in improper health behaviors, which brings inestimable harm to the public's physical and mental health. Thus, it is of great practical significance to study the identification methods of network pseudohealth information to prevent the spread of pseudohealth information and maintain social stability.

2. Related Works

Internet pseudohealth information mostly belongs to the nature of rumors, which have the characteristics of rapid spread, wide influence range, and great social harm. It often induces a wide range of network public opinions or public health events and attracts widespread attention. At present, the research on pseudohealth information identification mainly focuses on the following three aspects: (1) The "select instance" (or sliding window) classification method. For example, Molinaro and Greco proposed a two-stage instance selection algorithm, which is divided into two stages: concept detection and retraining. If the semantics of class health are detected, the algorithm will automatically update the classifier and find classification labels in class health information data for classification [1]. Han et al. proposed the sliding window algorithm, which can deal with the attribute classification problem of network pseudohealth information [2]. Hoens et al. proposed an support vector machine (SVM) model for detecting network pseudohealth information, and the classification of network pseudohealth information was realized by updating the weight allocation of instances [3]. (2) The batch classification method. For example, Sutskever et al. proposed the information batch processing model, which realizes batch processing of class health information by constantly updating the classifier, thus realizing the classification of pseudohealth information [4]. Rodriguez and Laio proposed an integrated model based on time limits, which can preliminarily compare and distinguish pseudohealth information and health information in the network [5]. (3) The classification method of online learning. For example, the pseudohealth information network online learning combination model proposed by Brzezinski and Stefanowski is composed of network online classifiers. Since the number of classifiers is usually fixed, as a result, the weighted sum update is also fixed [6]. Shi et al. proposed an online incremental algorithm to deal with the classification of network pseudohealth information. Due to the narrow value of online increments, which leads to poor fault tolerance [7], Eskandari and Javidi adopted the network online learning method to classify pseudohealth information through centralized processing, but its classification accuracy was relatively low, and the classification effect was also poor [8].

In previous related studies, scholars have proposed a variety of classification algorithms for the identification of network pseudohealth information, including the combination model of different algorithms. These algorithms and models have good recognition effects on pseudohealth information with obvious identification of information sources and text semantic tags. However, it is difficult to identify pseudohealth information with unclear information sources and unclear semantic tags in the network and is also difficult to detect and classify. In the previous research on pseudohealth information identification, whether it is "select instance" (or sliding window) classification method, batch classification method, or the online learning classification method, each has its own advantages and disadvantages. Although pseudohealth information can be classified from different aspects, the existing methods are mainly single classifiers or batch processing, which result in either the classification cannot be effective or the recognition accuracy not being high. Through the research on pseudohealth information, this paper aims to help people distinguish pseudohealth information and improve their health information literacy, thus fundamentally improving the quality of network health information and purifying the network health information environment. Based on this, this paper proposes an integrated combination of the latent Dirichlet allocation (LDA) algorithm and data partitioning and accurate update. By identifying network pseudohealth information by topic, class tag blocks are accurately updated and integrated with data blocks to effectively identify and classify pseudohealth information.

3. Research Methodology

3.1. Concept Definition. The combination algorithm proposed in this paper to identify the problem of network pseudohealth information, the core of which is to divide the

dataset corresponding to network pseudohealth information into "granularity" blocks according to its class label properties. To detect the minimum information unit attribute contained in the dataset, the dataset is continuously updated in blocks according to the category of information attribute contained in the minimum information unit and is reintegrated and classified according to the category of data blocks, to effectively identify pseudohealth information. The concepts involved in this combined algorithm are as follows.

Definition 1. Pseudohealth information (semantic definition). The so-called pseudohealth information refers to misleading others to follow blindly or accept false publicity in a misleading and deceptive way in the name of health to realize the personal interests of producers and broadcasters and has been falsified.

In summary, pseudohealth information usually appears in the external form of health information. It takes advantage of people's demand for health information and uses false, deceptive, misleading, and other ways and means to spread and advocate unscientific, false content to achieve personal purposes, and the information has been falsified. The semantics of pseudohealth information deviate from the information title and semantic label and have a conceptual drift with the original meaning. According to this, pseudohealth information can be defined in terms of information from the perspective of information dissemination, and its information definition is shown in Definition 2.

Definition 2. Pseudohealth information (information definition). Class health information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$, where x is the attribute value and y is the vector of the class label, decomposes its joint probability $P(x, y)$ into $P(x, y) = P(x)P(y|x)$. If the prior probability $P(x)$ and conditional probability $P(y|x)$ of the sample in the class health information dataset change, semantic concept drift occurs in the class health information dataset S: during semantic concept drift, if $P(x)$ does not change and $P(y|x)$ changes, it belongs to the concept drift of the conditional change class; that is, the class health information is determined as true health information; if $P(x)$ and $P(y|x)$ change, it belongs to the concept drift of feature change; that is, similar health information is false health information; that is, it is determined as false health information.

Generally, health information refers to similar health information in which the attributes or labels contained in the information dataset have not changed, but their external representations or conditions have changed over a period of time; however, pseudohealth information refers to those that appear as "health" and have a relatively stable feature distribution. However, the class health information is changed or deviates from the class label corresponding to the "health" eigenvector.

Definition 3. Data block. If the information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$ is divided into sequences arranged in sequence $z_1, z_2, \ldots, z_i, \cdots, z_n, \ldots$, each sequence contains a data record or several logical markers; if each sequence $z_i = (x, y)$ consists of eigenvectors $x \in X$ and class labels $y \in Y$, sequence elements z_i are called data blocks.

Definition 4. Data block integration. If the information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$ is divided into data blocks $z_1, z_2, \ldots, z_i, \cdots, z_n, \ldots$ with uniform size, each type of information data block contains d data blocks; for each newly added block z_j, the weight of the classifier $C_i \in \varepsilon$ is weighted by the weighting function $Q(\cdot)$. The weighting function $Q(\cdot)$ depends on the classification accuracy of the classifier. If the size of the data block set is set to k and does not exceed the limit, z_j is classified and added to a data block set of a certain type; if a data block set is a full set and the weight of the newly added data block is greater than that of the remaining data blocks, the newly added data block replaces the weakest block in the original set, and this process is called data block integration.

3.2. Algorithm Design

3.2.1. Algorithm Idea. The combination algorithm proposed in this paper blocks the data of the health-like dataset; that is, based on the class labels in the health-like dataset, topic recognition, information dataset partitioning, data block classification integration, and semantic offset detection are involved in the LDA model, Algorithm 1, SVM model, and Algorithm 2. The logical framework of the combined model is shown in Figure 1.

3.2.2. LDA Model. LDA was proposed by David Blei, Andrew Ng, and Michael I. Jordan in 2003. It is mainly used for document-topic generation and contains three levels of structure: document, topic, and word. Therefore, it is also called the probability model of the three-layer shellfish leaf stage [9]. As soon as the LDA model was proposed, it attracted the attention of scholars, especially in the field of semantic mining, which can greatly reduce the representation dimension of the text, thus making the model widely used [10, 11]. Additionally, as a typical representative unsupervised model, the LDA model has the advantage that the number of topics can be determined as long as important input parameters in the model are determined; therefore, the algorithm process is greatly simplified [12]. Based on this, when determining the optimal value of the number of document topics, this paper selects perplexity as an index to evaluate the pros and cons of the model, and its calculation equation is as follows:

$$\text{Perplexity}(D) = \exp\left[\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right]. \quad (1)$$

In the equation, M is the number of documents, D is the set of words in the document, w_d is the word, N_d is the number of words, and $p(w_d)$ is the probability of words in the document.

According to the statistical results, users who have published more Weibo information basically do not have the behavior of spreading fake health information, and their user credibility can be measured by the number of fans, the

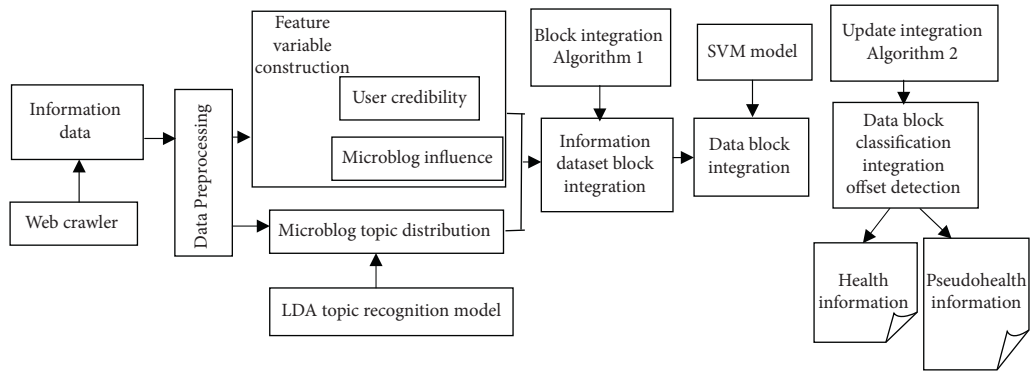


Figure 1: Logical framework of combined models.

number of followers, and the ratio; for those users who observe more but have fewer fans, the credibility is relatively low, and their fans are often the Internet Water Army. These users are most likely to be publishers or sources of a large amount of pseudohealth information. They publish or disseminate pseudohealth information through various network social platforms, such as Weibo and WeChat. Therefore, user credibility Reliability (u) can be defined by

$$\text{Reliability}(u) = \ln(\text{follower} - \text{following} + \text{num}) + \text{verify}. \tag{2}$$

In the equation, follower, following, and num are the number of fans, the number of followers, and the number of Weibo posts, respectively, after z − score standardization.Reliability (u) is an important basis for measuring the user credibility. The larger the value of Reliability (u) is, the higher the user credibility is.

For users, the number of fans, the number of Weibo forwards, the number of comments, and the number of praises are the basis for evaluating their influence. Generally, the more fans a user has, the greater the probability that the microblog posted by the user will be seen and spread by others and the more the corresponding forwarding, comments, and praise are. Regardless of what kind of Weibo operation behaviors fans perform, they all focus on the content published by users. Therefore, the influence of users' Weibo Influence (t) can be defined according to the following equation:

$$\text{Influence}(t) = \ln(\text{follower} + \text{repost} + \text{comment}) + \text{like}. \tag{3}$$

In the equation, follower, repost, comment, and like are the number of fans, forwarding, comments, and praise after z − score standardization. Influence (t) is an important indicator for evaluating the influence of users' microblogs. The greater the value of Influence (t) is, the greater the influence of user microblogs is.

3.2.3. Data Block Update Integration Algorithm

(1) Dataset Partitioning Algorithm. The identification of network pseudohealth information determines the essence of information semantics according to the deviation degree between the target class label and semantic ontology. If the semantic concept in information dataset S^t is replaced by S^{t+1} and the type of deviation is a subversive deviation, the "health" information content contained in the information semantics is replaced by pseudohealth information, its information semantic ontology has undergone fundamental changes, and the semantic ontology of network pseudohealth information belongs to this category. According to this principle, the information dataset S is now divided into data block streams $z_1, z_2, \ldots, z_i, \ldots, z_n, \ldots$, and each data block contains one record or several logical records. Classifier C_i is constructed, and the newly added data block Z_j is empowered. The classification performance of classifier C_i is determined by the weighted function $Q(\cdot)$. In the process of information dataset partitioning, if a certain type of data block set is not a full set, data block Z_j is added to this type of set; if a set of data blocks is a full set and the weight of block Z_j is greater than that of any block, the weakest block is replaced. The block integration algorithm of the dataset is shown in Algorithm 1.

(2) Data Block Set Classification Integration SVM Model. SVM is a typical representative binary classification model that is superior in classification generalization ability; therefore, it has been widely used in the field of information and data classification [13, 14]. In this paper, when identifying network pseudohealth information, the SVM model is adopted to integrate and classify the data block set to transform the instance sample dataset into the problem of solving convex quadratic programming. Then, the best classification hyperplane of the sample space is obtained. The classification hyperplane equation is as follows:

$$\omega^T \cdot x + b = 0. \tag{4}$$

In the equation, $\omega = (\omega_1, \omega_2, \ldots, \omega_7)$ is the normal vector, which determines the direction of the hyperplane; b is the displacement item, which determines the distance between the hyperplane and the origin; and $x = (f_1, f_2, \ldots, f_7)$ is the eigenvector of the sample point. The distance of the hyperplane is a controllable factor that makes the distance between two types of sample points and the classification hyperplane reach the optimal size based on the requirement of classification accuracy [15]. In addition,

the SVM model has good fault tolerance in the training process, and the optimal solution form of its optimal classification hyperplane equation is as follows:

$$\min_{\omega,b} \frac{\|\omega\|^2}{2} + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \begin{cases} T_i\left(\omega^T \cdot x_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ i = 1, 2, \ldots, N. \end{cases} \tag{5}$$

In the equation, x_i is the eigenvector of the i-th sample point, ξ_i is the relaxation variable of the i-th sample point, T_i is the category label of the i-th sample point, N is the number of training samples, and C is the penalty coefficient. The classification performance of the SVM model is determined by its kernel function. Choosing different kernel functions will lead to great differences in classification accuracy. At present, the kernel functions commonly used in SVM models include linearity, polynomial, and radial basis function (RBF) [16]. Since the classification accuracy of the RBF kernel function is much higher than that of other kernel functions and is suitable for situations where the number of features is less than or equal to the number of samples [17, 18], this paper chooses the RBF kernel function, as shown in the following equation:

$$K(x, z) = \exp\left(-\gamma\|x - z\|^2\right). \tag{6}$$

When the SVM model is classified and trained, the penalty coefficients C and γ in the RBF kernel function need to be determined in advance, the fault tolerance of the model is controlled by the former, and there is a negative correlation between the two; that is, the larger the penalty coefficient C is, the smaller the fault tolerance is. When C is too high, the overfitting phenomenon occurs [19, 20]; however, when C is small, to a certain extent, the classification accuracy of the model will be reduced accordingly. In other words, the parameter γ in the RBF kernel function affects the distribution of sample points mapped to high-dimensional space and exerts an influence on the penalty coefficient C, thus making the SVM model have high classification integration accuracy.

(3) Semantic Offset Detection Algorithm. The semantic changes in network pseudohealth information are very complex. Existing studies use online weighted and incremental classification methods to detect the changes in target semantics in network pseudohealth information, but data block integration is much more complicated than incremental classification, and the existing semantic offset detection algorithms have defects. To compensate for this defect, this paper adopts the semantic offset detection algorithm. The principle of this algorithm is that each data block contains one record or multiple logical markers, the data block set classified by the SVM model needs to be batch processed, and the candidate classifier corresponding to the integration component of the data block set is triggered for a classification check. If the current data block set is correctly classified, the original data block set of classification integration can be kept unchanged; if the fault tolerance of the current data block classification integration is poor or the classification accuracy is low, the integration component is reweighted, and the class tags in the data block are redetected to improve the classification accuracy of the classifier to effectively detect the target semantic attributes. Therefore, the semantic offset detection algorithm is shown in Algorithm 2.

4. Experimental Results and Analysis

4.1. Instance Data Acquisition. In this paper, crawler software is used to crawl experimental data, and pseudohealth information published by the publicity section of the Sina Weibo community management center is used as a reference. This pseudohealth information is reported due to false information and has been clearly confirmed by the government as pseudohealth information. Due to the spread of various pseudohealth information during the new coronavirus epidemic, the pseudohealth information in Sina Weibo is considerable. This paper crawled pseudohealth information from the API of Sina Weibo from February 1 to March 31, 2020, and randomly collected 1,183 pseudohealth information points. Among them, 759 of the original Weibo have more than 100 comments. The content of each Weibo was marked, counted, and sorted by its number of forwards, comments, and praises, and the experimental case dataset was constructed together with user information and the number of followers and fans.

To prevent the classifier from dividing all experimental data into health information, we added a manual verification step and selected some microblogs with comments greater than 100 and text, not pure symbols, and length greater than 10. The classification basis was obtained by means of manual verification technology and compared with health information. A total of 368 pieces of health information data were obtained through layer-by-layer screening, with more than 96.43 million comment texts. Based on the characteristics of comment anomaly parameters and SVM model parameters determined by the algorithm, this paper manually labeled the collected instance datasets. The selected instance dataset includes 359 pieces of pseudohealth information and 268 pieces of health information. When verifying the pseudohealth information recognition model, we made full use of the remaining 100 pieces of pseudohealth information and 100 pieces of health information to conduct precision comparison training experiments. The dataset composition of the experimental examples is shown in Table 1.

4.2. LDA Topic Recognition and Preprocessing. According to the data variables given in Table 2, the LDA model was used to preprocess the instance dataset to mine the document-topic distribution characteristics of the pseudohealth information dataset; the variables listed in Table 2 are the characteristic indicators required for LDA model preprocessing, and the meaning of each variable corresponds to

Input: S: an instance information dataset divided into blocks d; K: the number of members of the dataset; B: an instance buffer with a size of d; $Q(\cdot)$: classification quality measurement.
Output: ε: the integration of the classifier weighted as k.
(1) Information data block do for all $Z_j \in S$
(2) According to Z_j and $Q(\cdot)$, candidate classifier C' is established and empowered;
(3) According to Z_j and $Q(\cdot)$, all classifiers C_i in set ε are empowered;
(4) if $|\varepsilon| < k$, then $\varepsilon \longleftarrow \varepsilon \cup \{C'\}$;
(5) Else if $\exists i$: $Q(C') > Q(C_i)$, then replace the blocks in the weakest set with C';
(6) Initialize C' with B;
(7) $B \longleftarrow \varnothing$;
(8) Calculate the error of all types $d \in \varepsilon$ to S;
(9) Run the command on all instances of $Z_j \in \varepsilon$;
(10) End if
(11) End for

Algorithm 1: Dataset block integration algorithm.

Input: S: instance information data flow, D: information semantic offset detector, k: number of integrated members, B: instance buffer with size d, $Q(\cdot)$: classification quality measurement, t: number of instances;
Output: ε: offset detector integration with 1 classifier and k-class weighted classification;
(1) For all $x^t \in S$ instance do
(2) Gradually replace D with x^t
(3) $B \longleftarrow B \cup \{x^t\}$
(4) if $|B| = d$ or the offset is detected, then;
(5) According to W and $Q(\cdot)$, candidate classifier C' is constructed and empowered;
(6) According to $W$and $Q(\cdot)$, the classifier $C_i$in integration ε is empowered;
(7) if $|\varepsilon| < k$, then $\varepsilon \longleftarrow \varepsilon \cup \{C'\}$;
(8) Else if $\exists i$: $Q(C') > Q(C_i)$, then replace the weakest block in the integration with C';
(9) Initialize D;
(10) $B \longleftarrow \phi$;
(11) End if
(12) End for

Algorithm 2: Semantic deviation detection algorithm.

Table 1: Composition of experimental microblog datasets.

Category	Keyword	Number
Pseudohealth information data from February to March in 2020	Resistance viruses	217
	Immunity viruses	123
	Infection viruses	119
Health information data from February to March in 2020	Viruses	368

Table 2: Data variables.

Features	Document-topic distribution	User characteristics			Weibo features			
Feature metrics	0 ... n	Authentication	Number of fans	Number of attentions	Posted Weibo	Number of forwards	Number of comments	Number of praises
Variables	p_{mo} ... P_{mn}	Verify	Follower	Following	Num	Repost	Comment	Like

its characteristic indicators, where verif y_i indicates whether user u_i's Weibo account has been authenticated for personal information. If it was authenticated, u_i is 1; otherwise, it is 0. The characteristic indicators of other variables were consistent with the variable characteristic indicators in the user credibility and perplexity equation.

The result of LDA model preprocessing is shown in Figure 2. In the figure, the horizontal axis is the number of

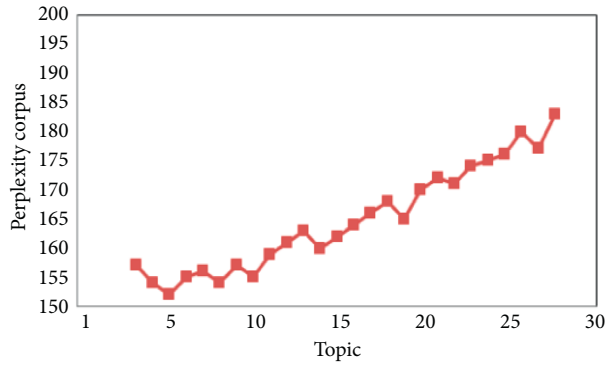


Figure 2: Perplexity-topic line chart.

Table 3: Distribution of subject words.

Topic 1	Alcohol	High temperature	Degree	Killing	Virus
Probability	0.043	0.039	0.037	0.015	0.009
Topic 2	Sterilizing fluid	Drinking	Virus	Killing	Effect
Probability	0.022	0.017	0.013	0.009	0.006
Topic 3	Mask	Multistory	Stopping	Stopping	Valid
Probability	0.075	0.046	0.033	0.028	0.025
Topic 4	Double *Coptis chinensis*	Inhibition	Virus	Mitigation	Treatment
Probability	0.049	0.039	0.027	0.021	0.021
Topic 5	5G	Spreading	Radiation	Carrying	Virus
Probability	0.036	0.023	0.023	0.008	0.005

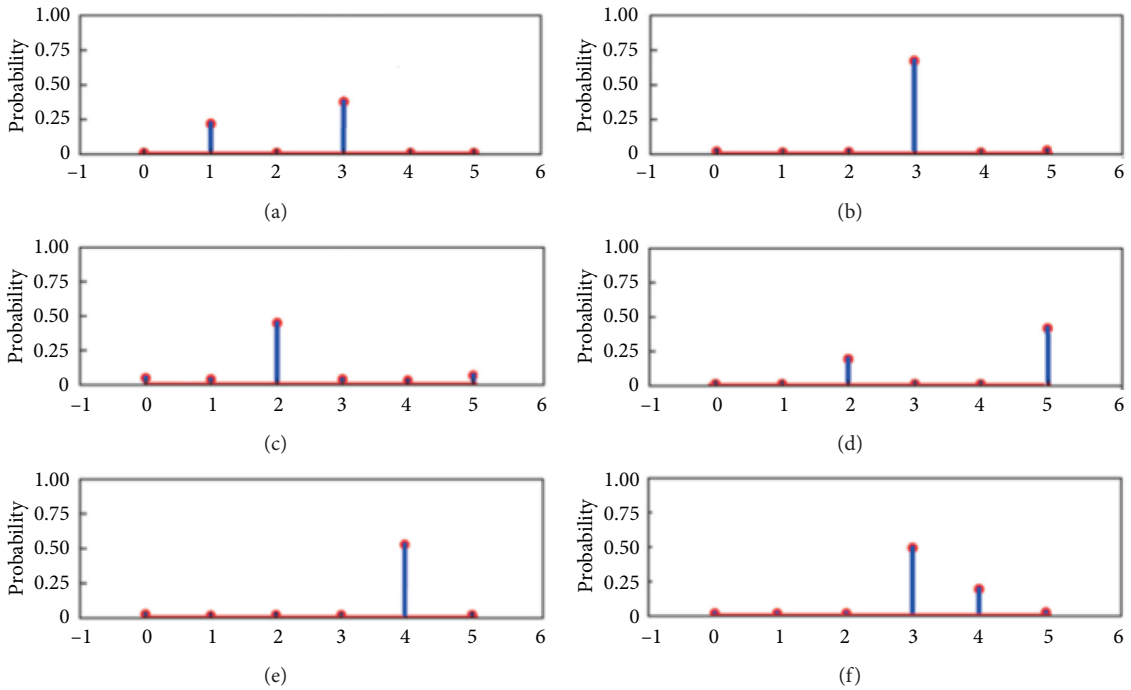


Figure 3: Document-topic distribution. (a) Topic, document 50. (b) Topic, document 61. (c) Topic, document 24. (d) Topic, document 39. (e) Topic, document 46. (f) Topic, document 45.

topics, the vertical axis is the perplexity, the polyline is 3 to 28, and the interval is 1. As seen in Figure 2, with the increasing number of subjects, the perplexity also continues to increase, but the rising track has certain volatility. When the number of subjects is 5, the perplexity reaches its lowest value. As the number of topics increases, the perplexity also

increases in a wave and reaches the maximum when the number of topics is 28. Based on the minimum principle of "perplexity + number of topics," 5 is selected as the topic parameter value of the LDA model.

After determining the optimal topic parameter value of the LDA model, the LDA model can be used to perform deep semantic training on the segmented instance dataset and then determine the distribution rules of "document-topic" and "topic-word" to determine the class labels or classification features of topics and words and prepare for the block and reintegration of instance datasets. The training results are shown in Table 3. As seen in Table 3, the results of LDA model training have obtained 5 topics. Now, the first 5 words are selected to represent each topic, and the probability of occurrence of each word is given.

Next, we randomly selected 6 documents as examples and show their "document-topic" distribution map to explore the probability of their themes and subject words. The specific results are shown in Figure 3. As seen in Figure 3, the probability of six document topics is different, but there is always a higher probability of one or two topics, while the probability of other topics is lower, which shows that the LDA model can divide the topic of microblog text well and provide a good foundation for the next step of this paper to block and integrate the microblog pseudohealth information instance dataset.

4.3. Integration of Dataset Partitioning and Classification

4.3.1. Block Experimental Datasets. The experimental dataset processed by the LDA model was cross-verified K times, and the instance dataset (S) was input. It was randomly divided into K subsets S' ($S' = \{S_1, S_2, \ldots, S_K\}$) with different sizes and mutual exclusion. In addition, S' was trained and tested K times; that is, in i iterations, subset S_i was retained as a test set, and the remaining subsets were used for training. The block efficiency is K iterations of training times divided by the total number of experiments. The$K$-fold cross-validation uses the classifier in Algorithm 1 to extract the weight of interactive information. The purpose of the cross-validation experiment is to verify the block efficiency and performance of Algorithm 1.

According to Algorithm 1, for a given instance dataset, if the attributes and class labels of the information text are obvious, the accuracy of the instance dataset is very high; if the attributes and class labels of the information text are vague or not clearly defined, the window algorithm (Algorithm 2) needs to be used to detect semantic deviation. In the process of grouping instance datasets, with the change in candidate classifier C', the classification discrimination boundary also changes. For all classifier C_i weights, the instance information dataset S is divided into data blocks of uneven size: $z_1, z_2, \ldots, z_i, \ldots, z_n, \ldots$; the candidate classifier C' is established according to Z_j and $Q(\cdot)$, and it is empowered accordingly so that the decision boundary of the instance dataset will not fall into the center point of one-dimensional, two-dimensional, and three-dimensional spherical Gaussian step by step, the cross-validation data blocks present Gaussian distribution, and the block discrimination boundary is composed of two hyperbolic surfaces. The block decision area is not simply connected but the area where the two elliptical contour lines formed by probability density are located, as shown in Figures 4(a) and 4(b).

In Figure 4(a), candidate classifier C' implements the partitioning of instance datasets according to the attributes of class labels. It uses all candidate classifiers C' in set ε to assign and update data blocks and creates k components to retain the original class labels of data blocks. The weight is updated based on the size of the instance buffer d to ensure that all data blocks have corresponding nonzero weights. In Figure 4(b), instance buffer d can not only retain the class tags of data blocks but also decide whether to replace the data blocks with the weakest class tags in the data block set according to classifier C_i. In addition, the data blocks with the weakest class labels can be removed or collected into the sets of other classes to effectively block instance datasets.

4.3.2. Data Block Classification and Integration. Vectorization is required for data block classification and integration. This paper uses the SVM model for classification and integration training, calls the libSVM tool, and adjusts the values of parameters C and γ to make the covariance matrix of the instance data block set distribution equal to obtain two n-dimensional spherical distribution information sets, namely, "health" and "pseudohealth" information data block classification integration datasets σ_1 and σ_2, where σ_1 and σ_2 are located on both sides of a $n-1$-dimensional normalized hyperplane. The hyperplane is the classification decision boundary of the two. The central line of the two n-dimensional spherical distributions formed by σ_1 and σ_2 is perpendicular to the hyperplane, as shown in Figure 5(a) and 5(b). In the process of classification integration, assuming $\exists i$: $Q(C') > Q(C_i)$, all classifiers C_i in the set ε are empowered according to Z_j and $Q(\cdot)$. If the errors of all types $d \in \varepsilon$ to S are equal, then all data blocks with $Z_j \in \varepsilon$ are classified and weighted. By measuring the Oushi distance from each data block to the ε-mean vector, the "minimum distance" of the boundary (hyperplane) is judged based on ε display and classification, classify and collect the weighted data blocks into the nearest dataset $\sigma_i (i = 1, 2)$, and the weakest data block in the set $\sigma_i (i = 1, 2)$ is replaced with C' to realize the preliminary classification and integration of data blocks, as shown in Figure 5(a).

The window algorithm (Algorithm 2) is different from other integrated classifiers. Its combination with the SVM model can continuously update and empower data blocks; therefore, data blocks are classified and integrated into the form of class labels, and the semantic deviation of data blocks can be effectively detected. The candidate classifier C' in Algorithm 2 and all the classifiers C_i in set ε determine the distance between the classification integration dataset and the superplane, which continuously updates the weight of the instance data block set d. The distance between the two types

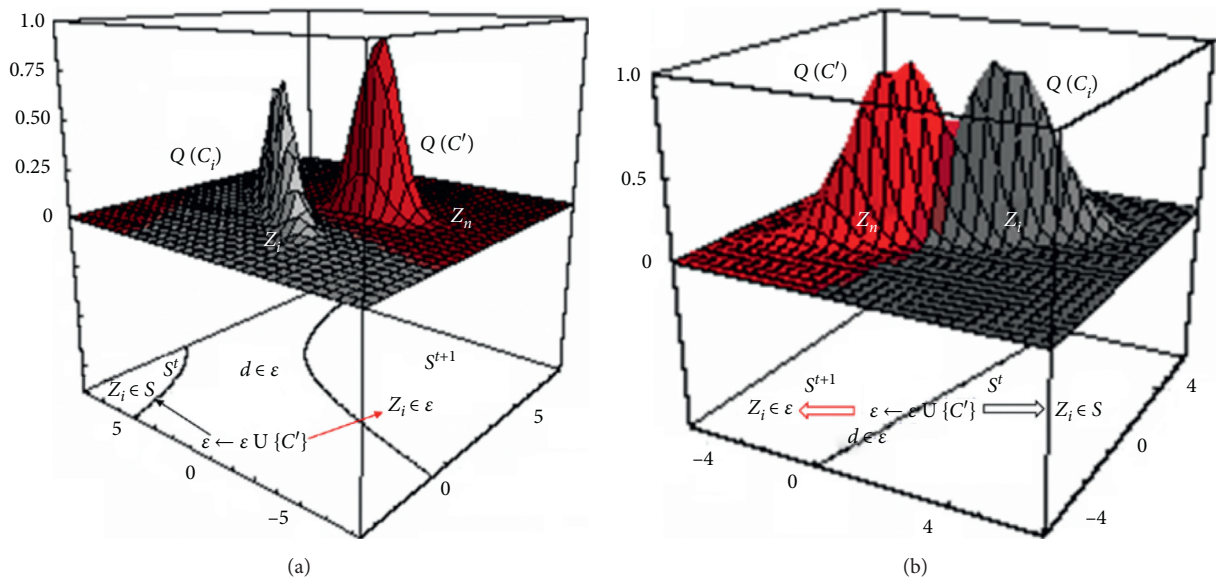


FIGURE 4: The process of instance dataset partitioning: (a) preliminary partitioning of candidate classifier C'; (b) weighting and updating of classifier C_i.

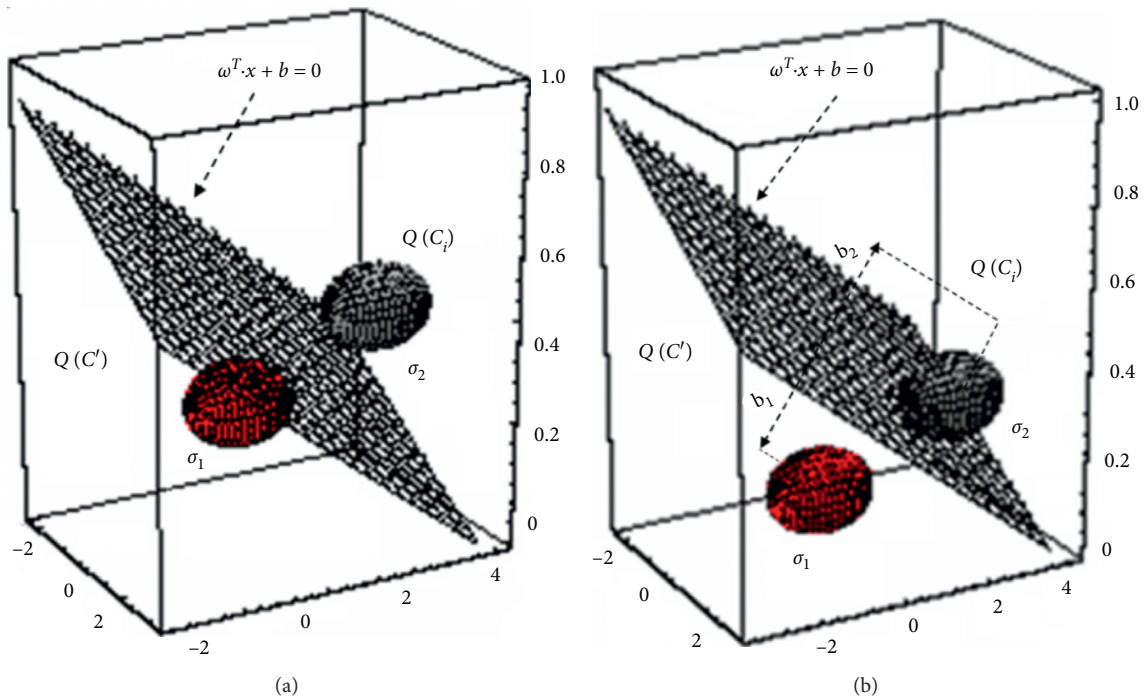


FIGURE 5: Classification integration process: (a) preliminary classification integration of data blocks; (b) precise classification integration of data blocks.

of datasets and the classification hyperplane is separated to the maximum through the best classification hyperplane in the SVM model (see equation (4)). At the same time, the slack variable ξ is introduced to improve the fault tolerance performance in the training process of SVM, and the sample points affected by the parameter γ in the RBF kernel function are mapped to the low-dimensional space to continuously correct the classification and integration efficiency so that the instance dataset can be accurately divided into "health" and "pseudohealth" information sets σ_1 and σ_2. The precise classification and integration process is shown in Figure 5(b).

4.4. Performance Evaluation of Classified Detection. To illustrate the advantages of the algorithm proposed in this paper, the logistic algorithm [21], decision tree (DT) [22],

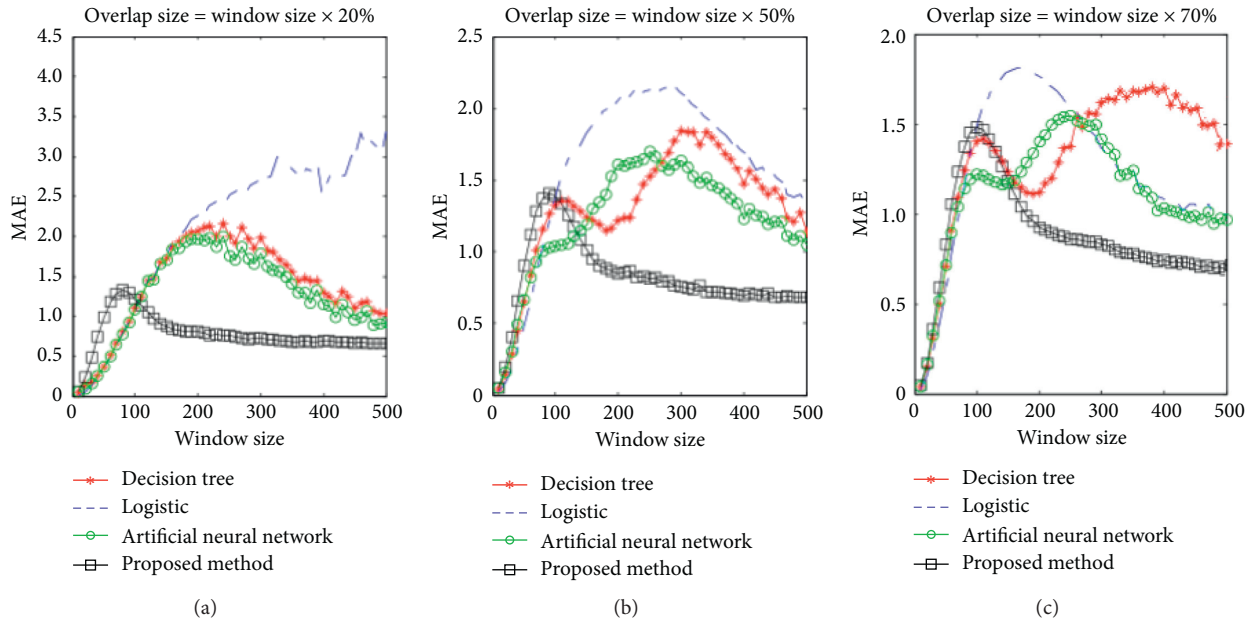


Figure 6: Detection performance of each classifier under different overlapping window size units: seconds.

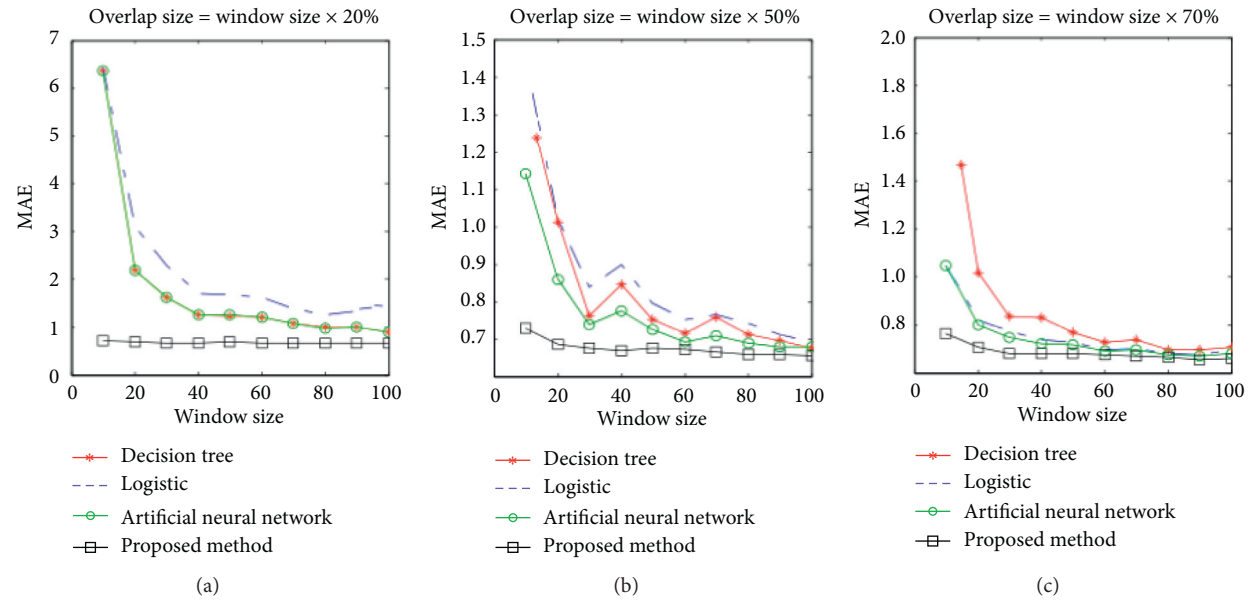


Figure 7: Detection performance of each classifier under different overlapping window sizes units: minutes.

and artificial neural network (ANN) [23] are now adopted for comparison. In addition, the classification accuracy among the four algorithms is tested. The classifiers of these four algorithms can update and classify the instance data block sets by using sliding windows in a free combination way. Therefore, the experimental data block set can be classified and integrated. Because the cross-validation strategy can overcome the overfitting of the classifier and enhance the generalization ability of the four algorithms, the classification accuracy of the four algorithms is compared by using the cross-validation strategy. The 10 instance subsets in this paper are randomly used for training to verify the classification accuracy of different models. The experimental results are shown in Figure 6.

As seen in Figure 6, the detection effects of the DT, logistic algorithm, and ANN are better than that of the method proposed in this paper within 0–100 seconds. However, the method in this paper is better than the other three algorithms in more than 100 seconds because the average absolute error (MAE) of the classification of the

Table 4: Classification accuracy of four classifiers units: %.

Data type	Classifier			
	DT	Logistic	ANN	Proposed method
Training sample	71.35	84.62	87.40	96.88
Sample test	76.29	86.07	91.18	98.73

method proposed in the paper is higher than that of the other three methods in 100 seconds; however, in more than 100 seconds, it is lower than that of the other three methods, and the three overlapping windows all have similar situations. To further illustrate this problem, the window unit is set to minutes, the sliding window size is 100 minutes, and the overlapping size is equal to 20%, 50%, and 70% of the window size. Four algorithms are used to detect the dataset of the example, and the detection effect is shown in Figure 7.

In Figure 7, the DT, logistic, and ANN algorithms can greatly reduce MAE by adjusting parameter settings and adopting supervised/semisupervised methods to increase the classification effect after 30 minutes. The algorithm in this paper can efficiently detect and classify instance datasets from the beginning, and its MAE value always fluctuates between 0.5 and 0.8. Therefore, whether in seconds or minutes, the algorithm in the paper is obviously superior to the DT, logistic algorithm, and ANN model.

The classification accuracy of the four algorithms is compared with the example dataset in the paper. The experimental results are shown in Table 4. As seen in Table 4, the classification accuracy of the four classifiers is quite different: the classification accuracy of the algorithm in the paper is the highest, the classification accuracy of the training sample is as high as 96.88%, the classification accuracy of the test sample is 98.73%, DT has the lowest classification accuracy, the classification accuracy of its training sample and test sample is 71.35% and 76.29%, respectively, the accuracy of the logistic algorithm and ANN is between the two, and the precision of ANN is slightly higher than that of the logistic algorithm.

5. Conclusions

The identification of network pseudohealth information is not only the frontier and focus in the field of news dissemination but also the focus and difficulty in the field of data mining. Although some scholars have studied this problem and proposed many recognition methods, the existing methods are mainly single classifiers or batch processing, which result in the fact that either the classification cannot be effective or the recognition accuracy is not being high. Based on the class tag attributes of network pseudohealth information datasets, the paper proposes a combination algorithm integrating data partitioning and classification update based on previous research results, integrates LDA topic recognition model, dataset partitioning algorithm, SVM data block classification integration model, semantic offset detection algorithm, and other methods, and adopts Web crawler technology to conduct simulation experiments based on the pseudohealth information of the Sina Weibo platform during the epidemic from February 1 to March 31, 2020. The simulation results show that the combination algorithm proposed in this paper has good superiority in both the subject recognition of pseudohealth information and the block and integration classification of instance datasets. Compared with DT, logistic algorithm, and ANN, the experimental results show that the classification integration accuracy of this method is higher than that of these three methods, which fully illustrates the reliability and practicability of the method in the paper. The identification of pseudohealth information in the future is of great significance for maintaining normal public health order and building a "Healthy China." Traditional mainstream media has high authority and influence. As a public tool of the society, media should perform its functions to serve the audience and the society and strengthen the check of fake health information to clarify its authenticity. At the same time, the media should also clarify the pseudohealth information that disturbs people in time to prevent the spread of pseudohealth information, which is also a way for the media to maintain their own image and authority. Therefore, we should not only pay attention to the problems existing in the dissemination of various kinds of information, but also make full use of technical means and tools to curb the further dissemination and influence of pseudohealth information.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the general program of the Natural Science Foundation of Shandong Province (no. ZR2019MG021) and the Key Projects of the National Statistical Scientific Research Plan (no. 2019LZ19). The research was also supported by the social science planning (dominant discipline) research project of Shandong Province (no. 19BYSJ19).

References

[1] C. Molinaro and S. Greco, "Polynomial time queries over inconsistent databases with functional dependencies and foreign keys," *Data & Knowledge Engineering*, vol. 69, no. 7, pp. 709–722, 2010.

[2] J. W. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*pp. 327–330, Morgan Kaufmann, Burlington, MA, USA, Third edition, 2011.

[3] T. R. Hoens, R. Polikar, and N. V. Chawla, "Learning from streaming data with concept drift and imbalance: an overview," *Progress in Artificial Intelligence*, vol. 1, no. 1, pp. 89–101, 2012.

[4] I. Sutskever, J. Martens, U. Dahl, and U. E. Hinton, "On the importance proceedings of initialization and momentum in deep learning," in *Proceedings of the International Conference on Machine Learning*, pp. 1139–1147, Atlanta, GA, USA, June 2013.

[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6, pp. 1492–1496, 2014.

[6] D. Brzezinski and J. Stefanowski, "Combining block-based and online methods in learning ensembles from concept drifting data streams," *Information Sciences*, vol. 265, no. 5, pp. 50–67, 2014.

[7] Y. Shi, F.-L. Chung, and S. Wang, "An improved TA-SVM method without matrix inversion and its fast implementation for nonstationary datasets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 2005–2018, 2015.

[8] S. Eskandari and M. M. Javidi, "Online streaming feature selection using rough sets," *International Journal of Approximate Reasoning*, vol. 69, no. 2, pp. 35–57, 2016.

[9] T. Gocken and M. Yaktubay, "Comparison of different clustering algorithms via genetic algorithm for VRPTW," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 574–585, 2019.

[10] J. Qu, Z. Ji, C. Lin, and H. Yu, "Fast consensus seeking on networks with antagonistic interactions," *Complexity*, vol. 2018, Article ID 7831317, 15 pages, 2018.

[11] D. Kurunathan, S. Shanmugathas, and K. Ashoka, "Analysis of relation between customer behavior and information technology market," *Journal of System and Management Sciences*, vol. 9, no. 1, pp. 87–104, 2019.

[12] Z. Han, X. Liu, and J. Kou, "Interdisciplinary subject recognition based on Rao-Stirling index and LDA model-a case study of nanotechnology," *Information Science*, vol. 38, no. 2, pp. 116–124, 2020.

[13] Y. Zheng, X. Hu, and J. Yin, "Health data fusion method based on multi-task support vector machine," *System Engineering Theory and Practice*, vol. 39, no. 2, pp. 418–428, 2019.

[14] Y. Yang, F. Zhang, and H. Xue, "A modal Fu Liye-support vector machine optimization method for abnormal data reconstruction of water intake monitoring," *Operations Research and Mamagement Science*, vol. 28, no. 2, pp. 52–59, 2019.

[15] X. Li, S. Wu, X. Li, H. Yuan, and D. Zhao, "Particle swarm optimization support vector machine model for machinery fault diagnoses in high-voltage circuit breakers," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 6, pp. 1–10, 2020.

[16] K. Bi and T. Qiu, "An intelligent SVM modeling process for crude oil properties prediction based on a hybrid GA-PSO method," *Chinese Journal of Chemical Engineering*, vol. 27, no. 8, pp. 1888–1894, 2019.

[17] Y. Lu, W. Wei, Y. Li, Z. Wu, and H. Jin, "The formation and evolution of interorganisational business networks in megaprojects: a case study of Chinese skyscrapers," *Complexity*, vol. 2020, Article ID 2727419, 17 pages, 2020.

[18] R. Goyat, G. Kumar, M. K. Rai, and R. Saha, "Implications of blockchain technology in supply chain management," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 92–103, 2019.

[19] S. Natalija and S. Dragan, "Accelerating multiple flow accumulation algorithm using MPI on a cluster of computers," *Studies in Informatics and Control*, vol. 29, no. 3, pp. 307–316, 2020.

[20] T. Saric, G. Simunovic, D. Vukelic, K. Simunovic, and R. Lujic, "Estimation of CNC grinding process parameters using different neural networks," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 6, pp. 1770–1775, 2018.

[21] M. Li and H. Xu, "Reliability window analysis of Gap zero gate based on Logistic model," *System Engineering Theory and Practice*, vol. 39, no. 2, pp. 531–538, 2019.

[22] T. Chen and L. Zhu, "Assessing the performance of Decision tree and neural network models in mapping soil properties," *Journal of Mountain Science*, vol. 16, no. 8, pp. 1883–1847, 2019.

[23] L. Macyszyn, C. Jedryczka, and R. Staniek, "Design and finite element analysis of novel two-stage magnetic precession gear," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 586–595, 2019.

Hindawi
Complexity
Volume 2020, Article ID 6694186, 13 pages
https://doi.org/10.1155/2020/6694186

Research Article

An Ensemble Learning Model for Short-Term Passenger Flow Prediction

Xiangping Wang, Lei Huang, Haifeng Huang, Baoyu Li, Ziyang Xia, and Jing Li

School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Jing Li; jingli@bjtu.edu.cn

Received 3 November 2020; Revised 16 November 2020; Accepted 8 December 2020; Published 19 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

In recent years, with the continuous improvement of urban public transportation capacity, citizens' travel has become more and more convenient, but there are still some potential problems, such as morning and evening peak congestion, imbalance between the supply and demand of vehicles and passenger flow, emergencies, and social local passenger flow surged due to special circumstances such as activities and inclement weather. If you want to properly guide the local passenger flow and make a reasonable deployment of operating buses, it is necessary to grasp the changing law of public transportation short-term passenger flow. This paper builds a short-term passenger flow prediction model for urban public transportation based on the idea of integrated learning. The goal is to use the integrated model to accurately predict the short-term passenger flow of urban public transportation, using Multivariable Linear Regression (MLR), K-Nearest Neighbor (KNN), eXtreme Gradient Boosting (XGBoost), and Gated Recurrent Unit (GRU) as the four seed models, and then use regression algorithm to integrate the model and predict the passenger flow, station boarding and landing, and cross-sectional passenger flow data of the typical representative line 428 in the "Huitian Area" of Beijing from January 1, 2020, to May 31, 2020. Finally, the prediction results of the submodels are compared with those of the integrated model to verify the superiority of the integrated model. The research results of this paper can enrich the short-term passenger flow forecasting system of urban public transportation and provide effective data support and scientific basis for the passenger flow, vehicle management, and dispatch of urban public transportation.

1. Introduction

According to the annual report on Beijing's Transport Development in 2020, by the end of 2019, the Beijing Public Transport Group has 28,271 buses and 1,620 routes in operation. The annual passenger volume of electric buses reached 3.564 billion, with an average daily passenger volume of 9.7377 million, providing great convenience for Beijing residents to travel, and it is the main undertaker of Beijing's surface public transportation.

In recent years, the characteristics of public transport network operation have become increasingly obvious; also, some potential problems gradually emerged, such as traffic jams during rush hours, traffic supply and demand not matching, a large number of passengers commuting security hidden danger in passenger flow gathering in a certain space, and some large activities, bad weather, and bus fault under the special operating environment urgent need for rapid evacuation etc. At the same time, with the increasing development of urban public transportation informatization, the Advanced Public Transportation System (APTS) has become an indispensable part of the construction of a "Smart City" as a result of the accumulation of massive public transport IC card data assets. At present, Beijing has built a bus GPS data acquisition system, line network management system, and other basic data such as BUS IC card, BUS GPS, and bus network. Thanks to the rapid development of computer technology, the methods of machine learning and deep learning display advantages such as high computational efficiency and strong data processing ability. Applications based on big data prediction technology, comprehensive and accurate projections for short bus traffic, are to promote

effective shuttle buses and other public traffic modes; to improve the utilization rate of public transport vehicles, optimization of vehicle dispatching, and the important measures to enhance the level of public transportation system management and operation; and also are the core of the realization of the intelligent transport system.

At present, the public transportation enterprises are in the process of actual operation, and the vehicle operation dispatching scheme formulation depends largely on historical experience. The forecasting ability for short-time bus passenger flow of each station, line and period is insufficient. It will inevitably lead to public transport vehicles not being able to get reasonable scheduling, affect the passengers, and impact on the effective running of the bus system. Therefore, it is of great importance to use big data situation analysis technology to accurately predict short-time bus passenger flow based on traffic IC card data and external weather data to analyze and master the transport demand and passenger flow rule of public transportation. Forecasting traffic demand is a core issue in any transportation system organization, and the future demand provided by predictive algorithms means that a reasonable supply can be planned in advance. Bus passenger flow related indicators reflect the passenger travel demand and regularity; can, for the operators in time according to the current system resource, adjust operation plans such as temporary or reduce extra trains and other transportation emergency cases combined effective disposal; and provide a scientific basis for narrowing the scope of the influence of the incident. As a result, it is necessary for public transportation to study the short-term passenger flow forecast, build higher prediction accuracy of the model, and obtain more reliable short-term passenger flow distribution, so as to solve the above problems effectively.

In this paper, the integrated learning method is introduced into the model of short-term bus passenger flow prediction, which significantly improves the accuracy of bus passenger flow prediction and provides a new modeling method for the quantitative research of public transportation, which has the dual significance of theoretical guidance and method innovation.

2. Literature Review

Short-term passenger flow prediction is an important part of the intelligent transportation system, which can be used to assist the adjustment of travel behavior, reduce passenger flow congestion, and improve the service quality of the transportation system. The evolution of the passenger flow prediction method is a process of continuous development and expansion, from the initial linear estimation model to the current model of machine learning and deep learning, gradually towards maturity. Generally speaking, short-term passenger flow prediction methods can be divided into two categories: parametric method and nonparametric method. The main difference between these two types of methods lies in the assumed functional dependence between independent variables and dependent variables [1].

In the traditional parametric methods, there are mainly Autoregressive Models (AR), Exponential Smoothing (ES) [2], Autoregressive Integrated Moving Average (ARIMA) model [3], and so on. ARIMA model is a linear combination of time-delay variables and error terms. Since the 1970s, the ARIMA model has become one of the commonly used parameter prediction methods and has been widely applied to the prediction of short-term traffic data such as traffic flow, travel time, and speed. Based on the historical passenger flow data collected by the urban rail transit automatic ticketing system, Cai et al. [4] used the ARIMA model to predict the passenger flow of Guangzhou metro. In addition, due to the seasonal and trend characteristics of passenger flow time series data, some researchers have applied Seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict passenger flow. In order to deal with the strong seasonal autocorrelation of the time series of passenger flow of Serbian railway, Milenković et al. [5] used the SARIMA model to predict the passenger flow of Serbian railway, which shows good prediction performance. Wang et al. [6] analyzed the rule of passenger flow in and out of Beijing subway station with time change, and the SARIMA model is used for modeling. The results show that the predicted results can accurately reflect the time change rule of passenger flow in and out of Beijing subway station. Because these parametric models assume linear relationships between variables with time delay, it is difficult to capture nonlinear relationships between variables, so the use of traditional parametric methods is limited [7, 8].

In order to better deal with the nonlinear characteristics of passenger flow data, the nonparametric method is introduced. Different from the parametric method, the nonparametric method is to establish the nonlinear relationship between input variables and output variables without prior knowledge. Therefore, it is more flexible and widely used in passenger flow prediction. Guo et al. [9] used 15 minutes of time interval summary of real traffic flow data compared, and the experiment shows that the adaptive Kalman filtering method can get a feasible prediction accuracy, especially under the condition of traffic high volatility, shows how to improve the adaptability of this method and, finally, puts forward the suggestions to improve the short-term traffic flow prediction performance. According to the characteristics of bus passenger flow and the law of changing with time, Deng et al. [10] proposed a prediction model of multicore least-squares support vector machine. The model fully considers the influence of historical data on bus passenger flow. Zhao et al. studied the passenger flow distribution in each period of the bus line by using the method of combining wavelet analysis and neural network and predicted the passenger flow of the short-time bus line, so as to realize the dynamic control and reasonable scheduling of the bus. Zhang and Yang [11] combined the main factors affecting passenger flow with the neural network self-learning method and established a subway passenger flow prediction model based on the neural network of spline weight function. Wang et al. [12] used a correlation analysis method to analyze the relationship between pedestrian flow and its influencing factors, extracted 11 important influencing

factors, and established a prediction model of pedestrian flow using the modular neural network. Among these nonparametric methods, neural networks are widely used because of their good adaptability, nonlinearity, and ability to map arbitrary functions [13, 14].

In the era of big data, the data processing capacity and prediction accuracy of the model have higher requirements. Researchers have made efforts to increase network density, and Hinton et al. [15] first proposed the concept of deep learning in 2006. Compared with the traditional neural network and other shallow learning models, deep learning is equivalent to a deeper neural network; that is, there are more hidden layers, which enable it to express more abstract and higher-level nonlinear features and more accurately capture the "deep" features of short-term passenger flow. Bai et al. [16], aiming at the short-term prediction of bus passenger flow, used the Deep Belief Network (DBN) to establish a prediction model. Compared with the classical parametric method and nonparametric method, this model shows a good predictive advantage. Li Bangpeng, respectively, used the convolutional neural network and the time-length neural network prediction model in deep learning to predict the future indoor spatial and temporal passenger flow distribution based on the real spatial and temporal passenger flow data and made a model comparison.

At present, integrated learning is a widely used method in machine learning, which integrates different learner sets so as to improve the accuracy of prediction [17]. In order to facilitate the collection, the mainstream of the current research is the design algorithm which promotes the weak learner to the strong learner and integrates multiple learners generated by the same algorithm. Freund and Schapire [18] proposed the Adaptive Boosting (AdaBoost) algorithm, which uses sequence sampling and has high operational efficiency and practical application value. The Bagging algorithm proposed by Breiman [19], which uses self-sampling to combine the base learner, was subsequently improved into Random Forest (RF) in 2001 [20] and had become the most classic algorithm in Bagging integration. In 1992, Wolpert [21] proposed the stacked generalization (stacked generalization) model, but the stacking algorithm only provides the integrated idea, for its selection of learning has certain subjectivity and then the selection of some scholars to study the certain research, such as Ledezma et al. [22] and Xu Huili, to use the genetic algorithm in the metamodel and the selection of the base model is optimized. The stacking algorithm has difficulty in obtaining the correct base learner assembly. Integrated learning, due to the combination of multiple learners, greatly improves the prediction accuracy and generally performs better than each component model, which benefits from the diversity among models, reduces the risk of using isolated models, and compensates for the shortcomings of each model [23, 24]. In addition, its models can solve many problems that a single model cannot solve. The passenger flow of urban public transport is dynamic and random, so it is difficult for a single model to fit its trend well, and integrated learning can better make up for this deficiency.

In conclusion, due to the complexity and randomness of bus passenger flow, as well as the higher requirements of big data on the data processing capacity and prediction accuracy of the prediction model, the use of traditional parameter methods and shallow neural network methods is limited. The application of deep learning, integrated learning, and other methods provides a new opportunity for accurately capturing the nonlinear characteristics of STW passenger flow and processing large quantities of multisource data.

3. Materials and Methods

3.1. Data Selection and Processing. This paper selects the card-swiping passenger volume, station boarding and landing volume, and section passenger volume data of the typical representative bus line 428 in the "Huitian area" from January 1, 2020, to May 31, 2020, for the key index prediction. The data source is the IC card data of Beijing Public Transport Group, with a total amount of about 107,000 pieces of data. Based on the basic analysis of the card-swiping data, it can be known that most of the bus operation time period is from 05:00 to 24:00, and the number of card-swiping times within 15 minutes during this time period is counted; that is, each indicator should get 76 data based on the granularity of 15 minutes a day. The processing of time series data first needs to be converted into a supervised sequence according to the set time step; that is, for certain data, it is considered that the data of its previous time step bar has obtained this data (time step is the number of time steps). In this process, the daily supervised sequence length is the original daily time series length minus the time step.

3.2. Analysis of the Key Indexes of Urban Bus Network Monitoring in "Huitian Area". This part monitors and analyzes the three key indicators related to the passenger volume of route 428, namely, the card-swiping passenger volume, station boarding and landing volume, and section passenger volume. The time frame is from January 1, 2020, to May 31, 2020.

Route 428 is metro Longze Station-Tiantong Beiyuan Station, including 32 stations. The operating mileage of the line is 13.9 km, the average one-way running time is 47.73 minutes, and the average running speed is 17.74 km/h. There are 20 vehicles in operation. There are 100 trains per day and 19 in peak hours. The average daily passenger throughput is 3,474.

3.2.1. Card-Swiping Passenger Volume. As shown in Figure 1, due to the impact of the epidemic, the passenger volume of card swiping during the Spring Festival and the epidemic prevention and control period after the festival was significantly lower than the normal situation before the festival, while the passenger volume of card swiping during the epidemic prevention and control period after the festival was generally low and slowly picked up, with a weekly increase.

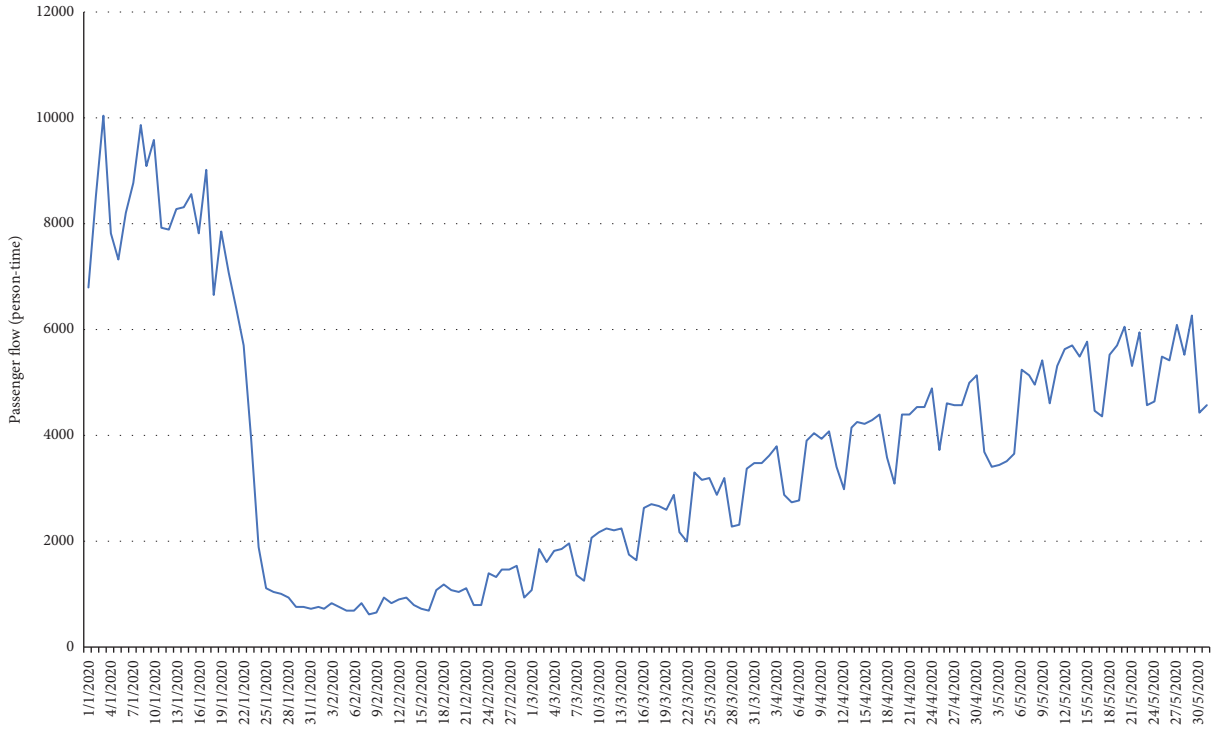


Figure 1: Passenger volume of the No. 428 bus.

3.2.2. Boarding and Landing Volume. The boarding and landing volume of bus number 428 is shown in Figure 2, in the direction of Metro Longze Station to Tiantong Beiyuan Station, the average daily volume for the largest station is 1036 (Banjieta Village North Station), and the average daily volume for the smallest station is 14 (the north gate of District 1, Harmony Garden). In the direction of Tiantong Beiyuan Station to Metro Longze Station, the average daily volume for the largest station is (Longjinyuan Area 4) and the minimum is 33 (Longxiyuan 3rd District Intersection West).

3.2.3. Sectional Passenger Volume. The average daily section passenger volume of bus number 428 is shown in Figure 3. The stations with the largest passenger volume in the direction of Longze Station and Tiantong Beiyuan Station are Banjieta Village North Station and Banjieta Village East Station. Tiantong Beiyuan Station–Metro Longze Station direction section passenger volume is the largest station for Xiaoxinzhuang East Station.

3.3. Model Selection. Bus passenger volume is affected by more external environment, and it is difficult for a single model to learn its complicated rules. Short-term prediction is essentially a question of time sequence, to the problem of the prediction which is usually not a model that can be applied to all scenarios, and integrated thinking is through a combination of several single models to reduce the risk of the error model, by giving full play to the information of the prediction results of each submodel to make up for the shortcoming of single model that the prediction error is large due to the influence of random factors, thus improving the prediction performance. This paper constructs four seed models of Multivariable Linear Regression (MLR), K-Nearest Neighbor (KNN), eXtreme Gradient Boosting (XGBoost), and Gated Recurrent Unit (GRU) and also constructs the regression integration model.

3.3.1. MLR. In this paper, we study the influence of many factors and so the selection of the most commonly used multiple linear regression, the simple model principle as shown in Figure 4.

3.3.2. KNN. KNN is a model based on distance. Figure 5 shows the algorithm principles of the classification model, according to the K value selection near the element, the element near the largest number of categories.

3.3.3. XGBoost. XGBoost is a boosting tree model based on ensemble learning boosting, which is based on regression tree. Once proposed, this method has been widely used in much research and many enterprises because of its high efficiency and accuracy. Some studies have shown that the prediction accuracy of this method can be comparable to the neural network and deep learning in dealing with time series problems.

3.3.4. GRU. GRU combines the forget gate and the input gate into one and mixes the cell state C and the hidden state. The final model is simpler than the standard LSTM, as shown in Figure 6.

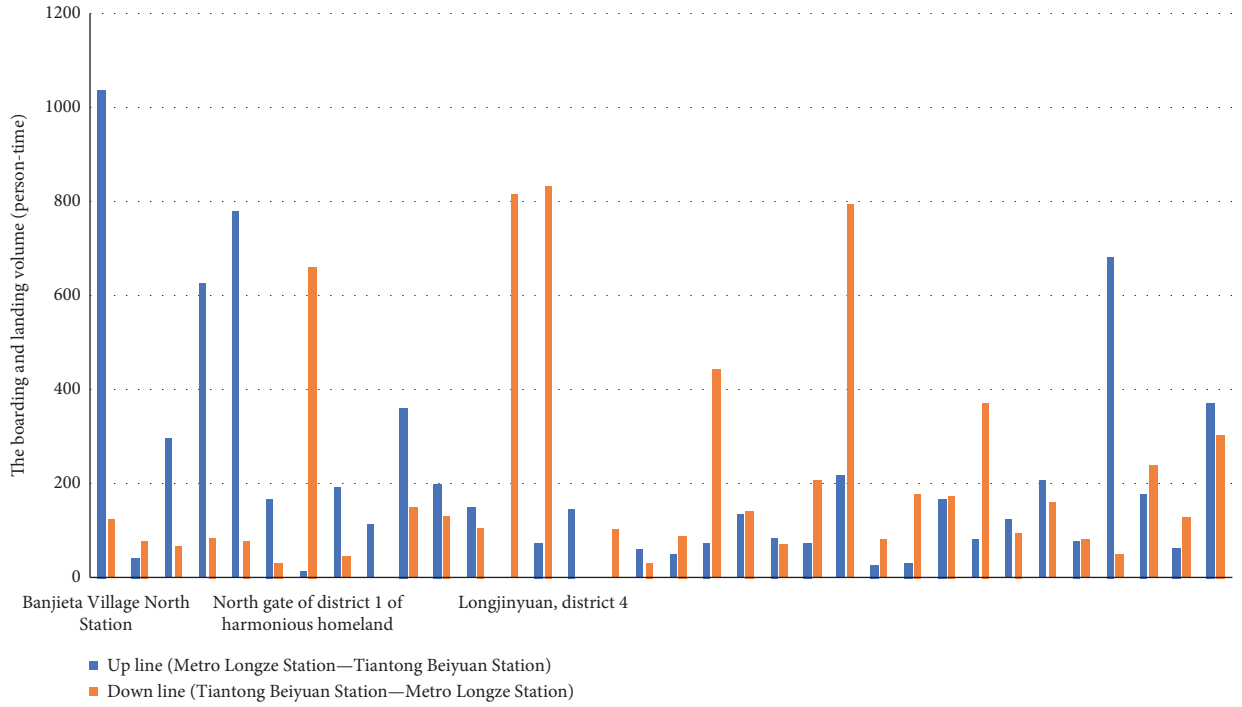


Figure 2: The boarding and landing volume.

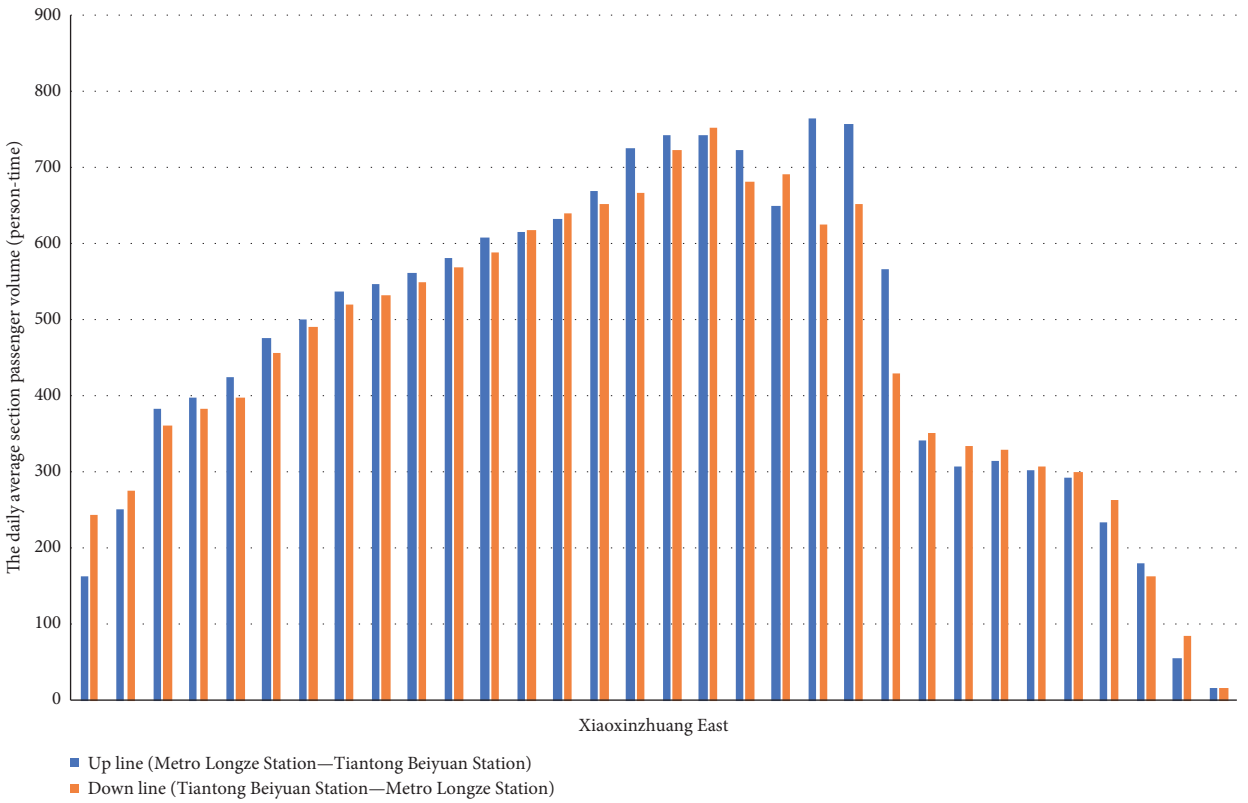


Figure 3: Daily average section passenger volume of the number 428 bus.

3.4. Build the Regression Integration Model. Integrated learning is an idea rather than a specific algorithm in machine learning. The core of this method is to combine multiple models called weak learners into a more accurate model. The integrated model uses different sampled data to train these weak learners continuously, adjusts the weak

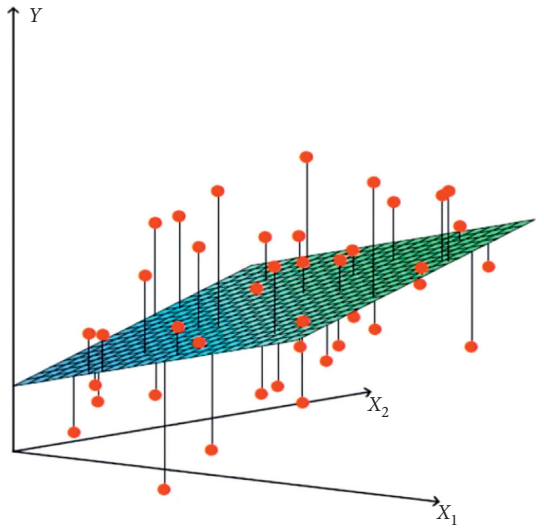


Figure 4: Schematic diagram of linear regression.

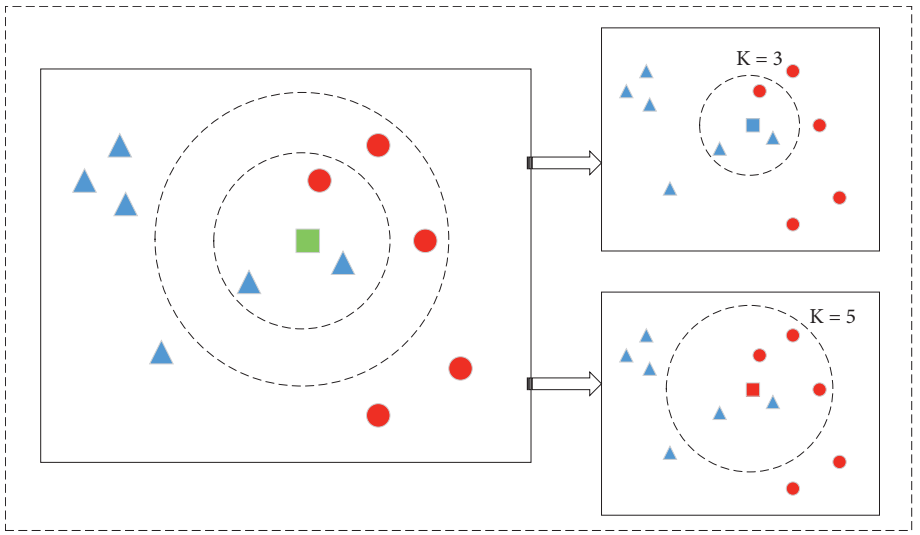


Figure 5: Schematic diagram of the KNN algorithm.

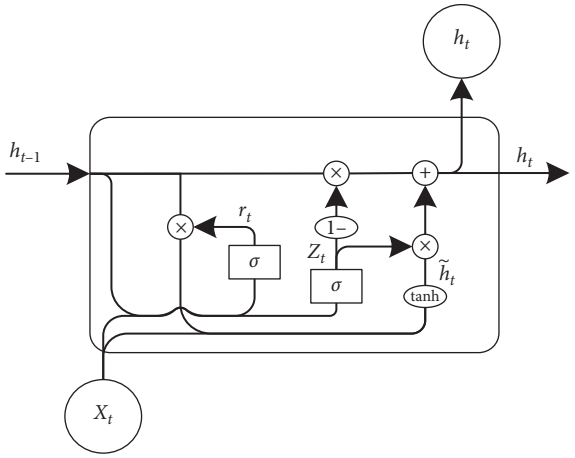


Figure 6: Internal structure of GRU.

learners through errors, and effectively combines the predicted results of the weak learners to a certain extent.

3.4.1. The Advantages of Integrated Learning. If the individual model is compared to a decision-maker, the integrated learning approach is equivalent to multiple decision-makers working together to make a decision. The advantages of ensemble learning are as follows: (1) overall, ensemble learning has a high accuracy rate; (2) the introduction of randomness makes it not easy to overfit, has good antinoise ability, is not sensitive to outliers of abnormal points, and can handle high-dimensional data without making feature selection; and (3) it can process both discrete data and continuous data. In addition, the data set does not need normalization, so the overall training speed is considerable.

3.4.2. The Regression Integration Model Based on GBDT. In this part, we combine the prediction results of the four-seed model with the regression model. This paper selected the regression model is Gradient Boosting Decision Tree (GBDT); this algorithm is based on the integration of learning. The passenger flow results predicted by each submodel are input into the GBDT model as an independent variable and the real value of passenger flow as a dependent variable for a new round of learning. Some nonlinear relations between the predicted results of the submodels and the real values can be learned through regression models, and the advantages of different submodels can be brought into play to make up for the disadvantages of different models. The model is shown in Figure 7.

3.5. Set Evaluation Index. In order to more comprehensively compare the different prediction results caused by the selection of different parameters in the same model, this paper selects Root Mean Square Error (RMSE) as the objective function of the optimization model and selects Mean Absolute Error (MAE) as the index of the evaluation model. Its definitions are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\mathrm{true}} - y_{\mathrm{pred}}\right)^2}, \tag{1}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{\mathrm{true}} - y_{\mathrm{pred}}\right|, \tag{2}$$

y_{true} represents the actual value, y_{pred} represents the predicted value, and N represents the predicted sample number. Both indicators reflect the size of the error between the predicted value and the actual value, but the former is more able to amplify the error, while the latter reflects the true error. The smaller the values of RMSE and MAE, the closer the predicted value to the actual value and the higher the prediction accuracy of the model.

4. Results

The passenger flow of swiping card and the boarding and landing volume reflects the passenger flow of a certain line or a certain station, and the passenger flow of section reflects the passenger flow between two adjacent stations on the line. The three indexes correspond to the essential basic data for optimizing the design of the route network and deploying vehicles in the public transport system, as well as the important basis for planning the bus dispatching frequency and considering whether to set interregional buses. Therefore, this paper selects three basic indicators of passenger flow—section passenger flow, card-swiping passenger flow, and boarding and landing volume for short-term prediction, providing the basis for rational planning of bus network, allocation of bus station facilities, and preparation of its operation plan.

4.1. The Prediction of Section Passenger Volume. This part selects the passenger volume data of the section with a grain size of 15 minutes from January 1, 2020, to May 31, 2020, in the upward direction of Xiaoxinzhuang East Station of number 428 bus in the "Huitian Area". Excluding the data not in the bus operation time, there are a total of 76 pieces of data in a day, with a total of 11,552 pieces of data. The time step was selected as a comparison of 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours. In other words, time step values were 1, 2, 4, 8, 12, and 24. In addition, the data ratio of the training set, verification set, and test set is 7 : 1 : 2, with 8086, 1156, and 2310 pieces of data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 1 and 2 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 8–13.

4.2. The Prediction of Card-Swiping Passenger Volume. This part selects the card-swiping passenger volume data with a grain size of 15 minutes from January 1, 2020, to May 31, 2020, in the upward direction of number 428 bus in the "Huitian Area." The time step was also selected to compare 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours. The data ratio of the training set, verification set, and test set was 6 : 2 : 2, with 8812, 2938, and 2938 pieces of data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 3 and 4 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 14–19.

4.3. The Prediction of Boarding and Landing Volume. In terms of the boarding and landing volume, 81,400 pieces of data have been collected from the North Station of Banjieta Village from the 15-minute ascending direction of bus number 428 from January 1, 2020, to May 31, 2020. In the same way, the data was converted into a supervised sequence according to the set time step; the time step was 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours, respectively, and the data ratio of the training set, verification set, and test set was 6 : 2 : 2, with 48,840, 16280, and pieces of 16280 data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 5 and 6 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 20–25.

5. Discussion

According to the "no free lunch" theorem in machine learning theory, there is no algorithm that can solve all problems perfectly. Many factors such as the size and structure of the data set will affect the final result. For specific data sets and actual needs, we should consider how to choose a suitable algorithm. This paper proposes a method for

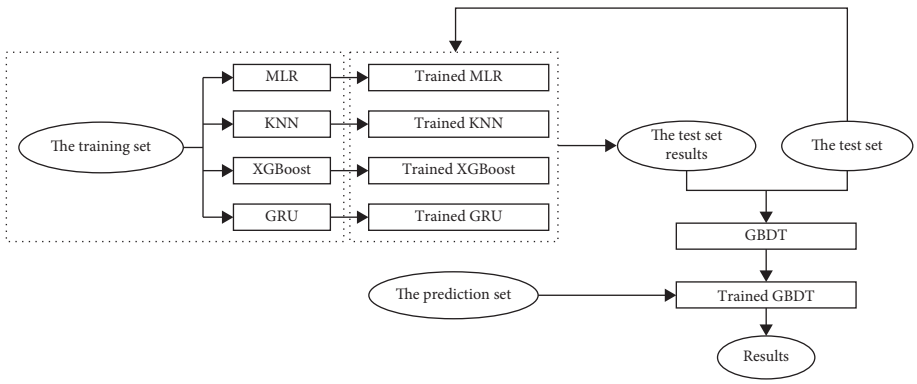


Figure 7: Ensemble flowchart based on the regression model.

Table 1: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	5.00	4.71	4.71	4.65	4.62	4.68
KNN	5.27	5.21	4.97	4.95	4.96	5.22
XGBoost	5.11	5.30	5.10	4.83	4.78	4.68
GRU	4.97	4.71	4.70	4.78	4.62	4.67
REG	4.94	4.61	4.70	4.62	4.57	4.67

Table 2: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	7.40	6.69	6.72	6.66	6.65	6.81
KNN	7.68	7.52	7.14	6.99	7.05	7.45
XGBoost	7.51	7.77	7.33	6.91	6.84	6.72
GRU	7.33	6.89	6.75	6.75	6.54	6.65
REG	7.30	6.64	6.71	6.59	6.50	6.64

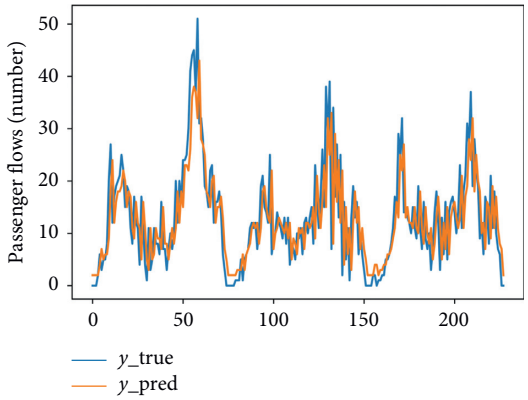


Figure 8: Regression integration prediction results when the time step is 15 min.

selecting the optimal model in regression prediction. The focus of this method is not the final specific model, but the selection process of the optimal model. Therefore, it is not

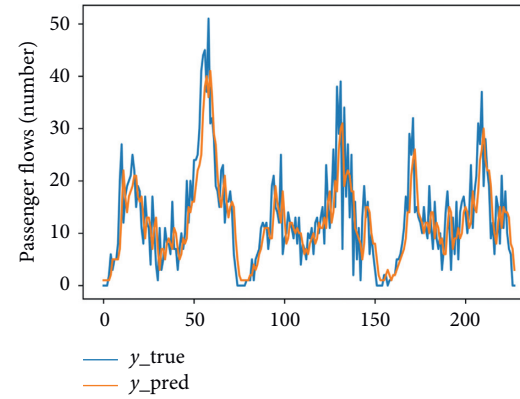


Figure 9: Regression integration prediction results when the time step is 30 min.

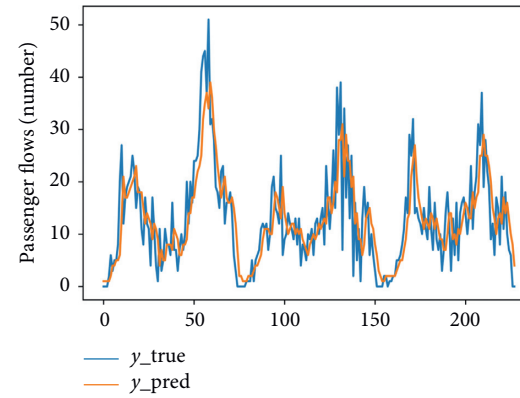


Figure 10: Regression integration prediction results when the time step is 2 h.

limited to being used on a given data set. This is exactly the innovation of this article. The integrated model selection part is the focus of this article. Realize regression prediction for boarding and landing volume, cross section passenger flow, and card-swiping passenger flow. The same regression integration algorithm can be used to predict all MAE and RMSE. The submodels can also be divided into several categories, and different algorithms can be selected for

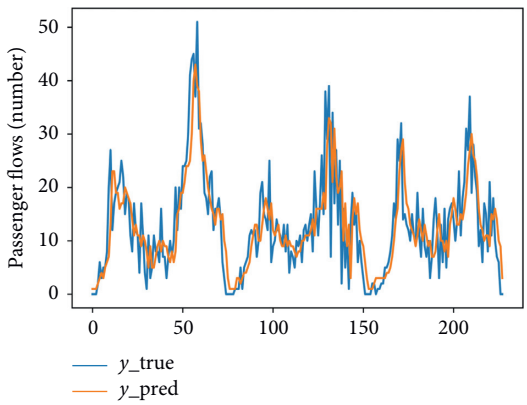


Figure 11: Regression integration prediction results when the time step is 2 h.

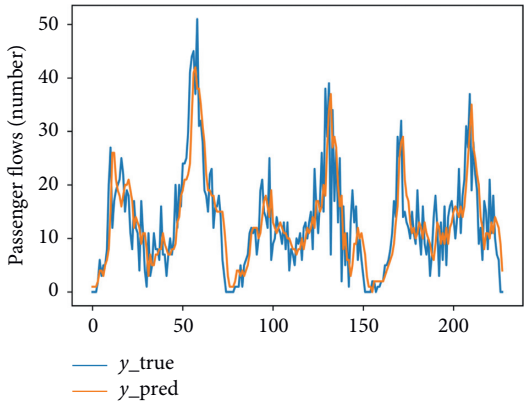


Figure 12: Regression integration prediction results when the time step is 3 h.

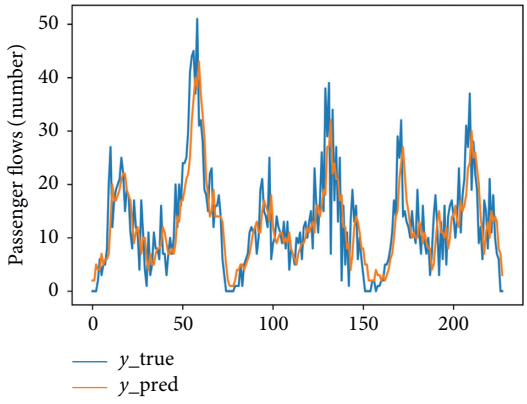


Figure 13: Regression integration prediction results when the time step is 6 h.

different types of indicator data sets. If different indicators are classified and predicted, there are problems of how to classify and which algorithm to choose. The method proposed in this paper is to use multiple algorithms to predict each index separately, select the optimal integrated model, and propose a comparative model to verify whether the selected

Table 3: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	17.87	17.67	16.85	16.93	17.53	20.74
KNN	19.84	18.84	18.93	21.37	18.24	18.71
XGBoost	19.91	20.47	18.72	18.75	17.45	17.51
GRU	18.05	18.04	16.81	19.39	18.34	18.64
REG	17.71	17.56	16.80	16.09	16.31	17.03

Table 4: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	17.87	17.67	16.85	16.93	17.44	19.42
KNN	19.84	18.84	18.93	21.37	17.82	18.09
XGBoost	19.91	20.47	18.72	18.75	18.43	18.47
GRU	18.05	18.04	16.81	19.39	17.25	17.53
REG	17.71	17.56	16.80	16.09	16.90	17.22

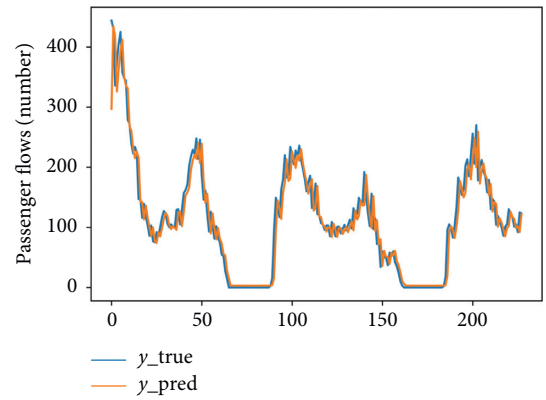


Figure 14: Regression integration prediction results when the time step is 15 min.

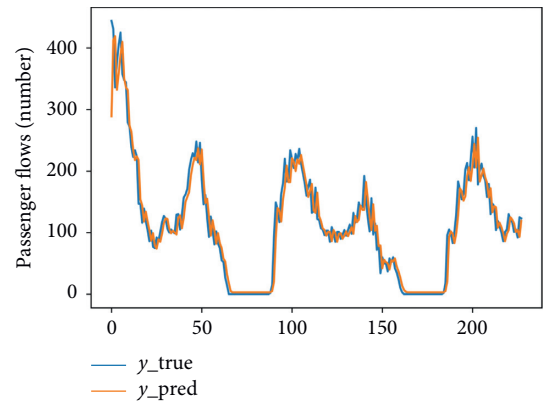


Figure 15: Regression integration prediction results when the time step is 30 min.

optimal integrated model performs best. In the empirical study, four machine learning algorithms of KNN, LR, XGBoost, and GRU were used to predict boarding and landing volume, cross

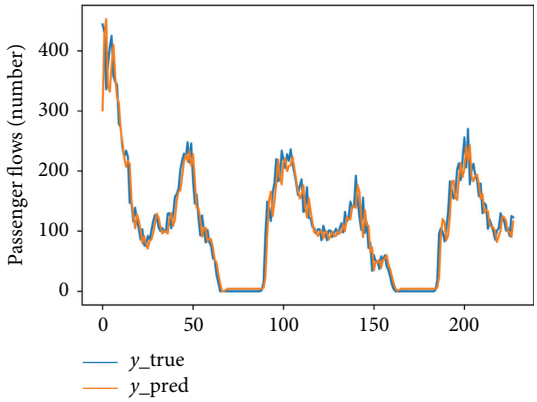


Figure 16: Regression integration prediction results when the time step is 1 h.

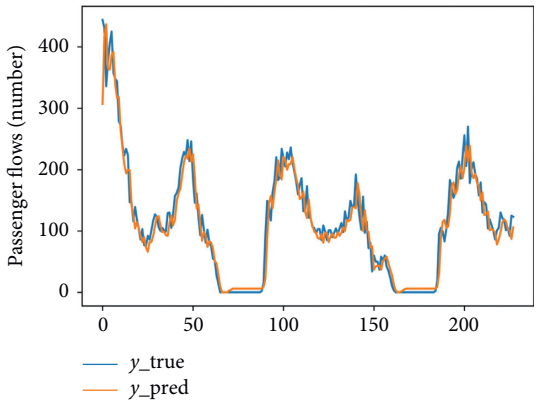


Figure 17: Regression integration prediction results when the time step is 2 h.

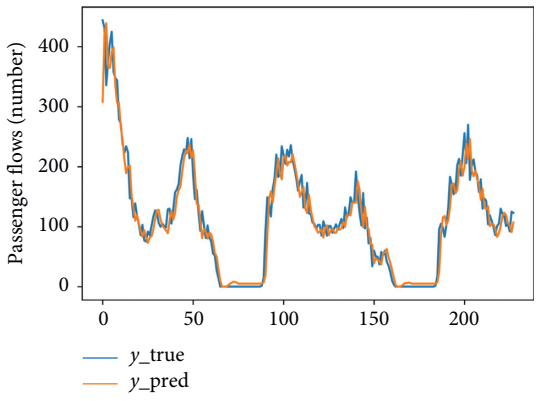


Figure 18: Regression integration prediction results when the time step is 3 h.

section passenger flow, and card-swiping passenger flow, respectively, finally comparing the prediction results of linear regression integration algorithms.

By comparison, in the cross section passenger flow prediction, the prediction results of LR and GRU at each step in the four submodels have lower MAE and RMSE values; the

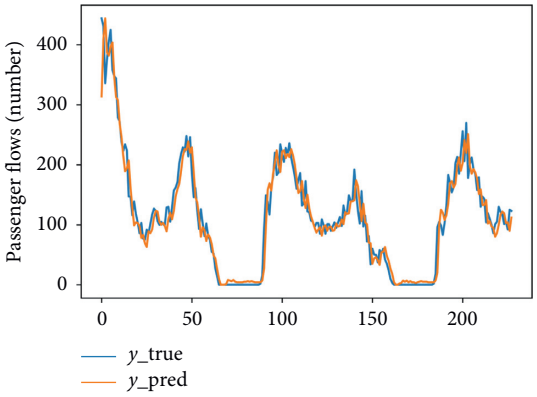


Figure 19: Regression integration prediction results when the time step is 6 h.

Table 5: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	5.79	5.19	5.04	5.05	5.05	5.04
KNN	5.95	5.65	5.44	5.05	4.96	5.00
XGBoost	5.71	5.37	5.41	5.06	4.91	4.86
GRU	5.70	5.04	4.94	4.73	4.82	4.72
REG	5.32	4.95	4.81	4.70	4.68	4.67

Table 6: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	9.51	8.56	8.46	8.36	8.33	8.31
KNN	9.86	9.70	9.12	8.53	8.28	8.32
XGBoost	9.41	9.47	9.18	8.37	7.91	7.81
GRU	9.43	8.56	8.40	8.10	7.80	7.80
REG	9.37	8.54	8.37	7.95	7.70	7.78

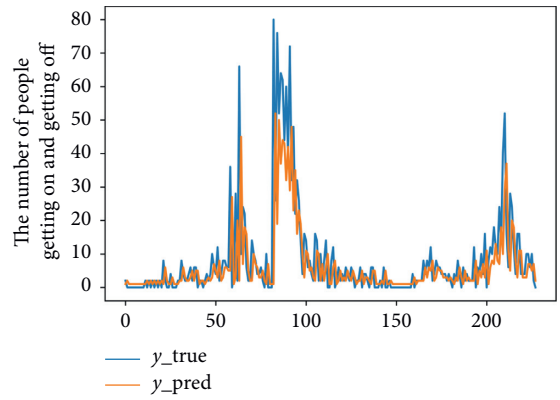


Figure 20: Regression integration prediction results when the time step is 15 min.

prediction results are more accurate; each submodel when the step length is 8 and 12 has relatively low MAE and RMSE values; and the results are more accurate. In comparison with

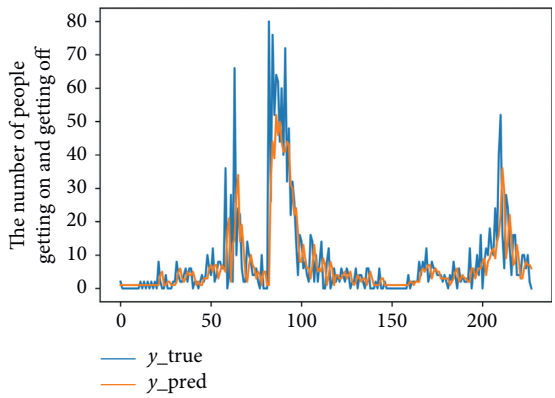


Figure 21: Regression integration prediction results when the time step is 30 min.

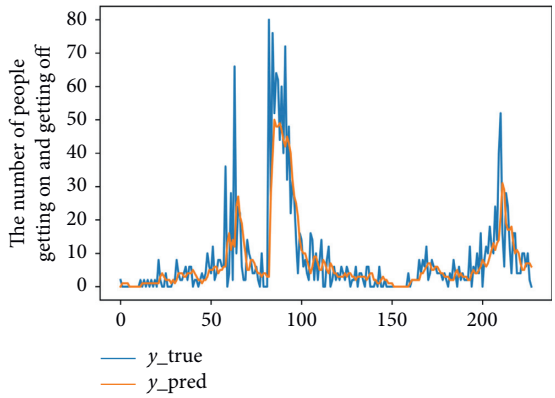


Figure 22: Regression integration prediction results when the time step is 1 h.

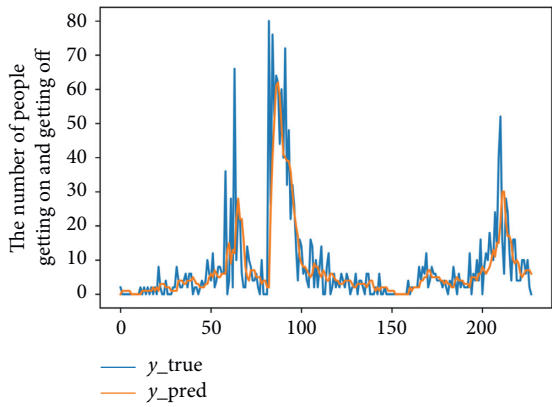


Figure 23: Regression integration prediction results when the time step is 2 h.

the prediction results of the regression integrated model, the integrated model has the lowest MAE and RMSE values at each step, indicating the result of using the regression integrated model to predict the most accurate. When the step size of the regression ensemble model is 12, the MAE value is 4.57 and the RMSE is 6.50, which are both the lowest values at each step size, indicating that the regression ensemble model has good prediction accuracy when the step size is 12.

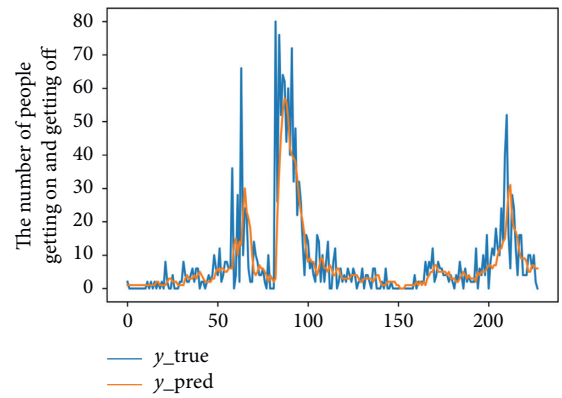


Figure 24: Regression integration prediction results when the time step is 3 h.

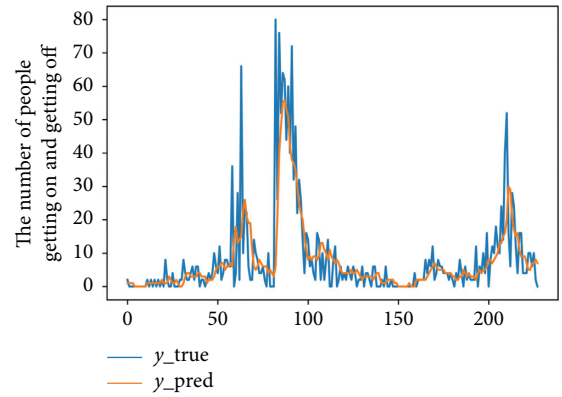


Figure 25: Regression integration prediction results when the time step is 6 h.

In the prediction of passenger flow by swiping cards, the prediction results of LR in the four submodels at each step have lower MAE and RMSE values, and the prediction results are more accurate. When the step is 4, each submodel has relatively low MAE and the RMSE value is more accurate. In comparison with the prediction results of the regression ensemble model, the ensemble model has the lowest MAE and RMSE values at each step, indicating that the prediction results of the regression ensemble model are the most accurate. When the step size of the regression ensemble model is 8, the MAE value is 16.09, and the RMSE is 16.09, both of which are the lowest values at each step size. It shows that the regression ensemble model has good prediction accuracy when the step size is 8.

In the prediction of landing volume, the prediction results of GRU under each step size in the four submodels have lower MAE and RMSE values, and the prediction results are more accurate. When the step size is 12, each submodel has relatively low MAE and RMSE. The results are more accurate. In comparison with the prediction results of

the regression integrated model, the integrated model has the lowest MAE and RMSE values at each step, indicating that the prediction results using the regression integrated model are the most accurate. When the step size of the regression ensemble model is 8, the MAE value is 4.68 and the RMSE is 7.70, which are the lowest values at each step size. It shows that the regression ensemble model has good prediction accuracy when the step size is 8.

6. Conclusions

The core of urban bus network operation management is to effectively allocate and use system resources according to changes in the bus network passenger flow, adjust operation strategies in time, and ensure that the bus network safely completes transportation service tasks. Short-term passenger flow prediction and analysis is the basis of operation management. It can provide a basis for emergency management and response and is also an important decision-making index for public transportation service level and system operation status evaluation. Short-term passenger flow prediction is an important decision data for urban public transportation operation and management, and its prediction accuracy will directly affect urban public transportation decision-making, adjusting the scientificity and accuracy of the operation plan.

This paper analyzes the operational monitoring data of 428, a typical line in the Huitian area, from the perspective of the urban public transport network in the Huitian area, including traffic capacity, as well as the boarding and landing volume and cross-sectional passenger flow of each station. At the same time, based on objective bus operation data, the lr, KNN, Xgboost, and GRU four-seed models and the regression integration model based on the four-seed model were used to predict three different passenger flow indicators. From the prediction results, it can be seen that the regression integration is compared with the other four submodels and the model has a higher degree of fit. For passenger flow prediction, the result of this integrated model has a high degree of credibility.

The reliability of the prediction results reflects the availability and effectiveness of the prediction methods and models to a certain extent and also ensures the availability of the final short-term passenger flow prediction results. According to the reliable prediction results, once the passenger flow prediction value is greater than the preset threshold, decision-makers can activate emergency management plans. Secondly, operational planning can be dynamically adjusted based on passenger flow fluctuations. Managers can effectively control short-term passenger flow changes, adjust network operation strategies in a timely manner, rationalize the use of public transportation resources, and reduce operating costs. At the same time, the result of the short-term passenger flow forecast is used as a positive feedback of the line network monitoring, which can assist the manager in obtaining more effective information from the daily bus line network monitoring, so as to improve the control and management of the bus line network.

Since the research in this paper focuses on the construction and verification of the basic model, there are still certain shortcomings and limitations. Based on these shortcomings and limitations, the following prospects and suggestions can be provided for future related work:

(1) The impact of traffic policies on individual travel characteristics is a long-term impact. At the same time, traffic data can accurately record the long-term travel activities of each individual; therefore, urban big data such as traffic card data is very suitable for analyzing the impact of changes in urban traffic policies, the influence of individual travel characteristics. In the later period, we can use the data over a long period of time to analyze the impact of urban traffic policy changes on individual travel characteristics from a longitudinal perspective.

(2) The addition of more source data and the improvement of richer individual attribute information: the addition of mobile phone data and other data including complete travel chain data can significantly improve the identification of passenger activity locations. This will improve the analysis of the generation mechanism of rail transit passenger flow and enrich individual attribute information more accurately. In addition, the data including the complete travel chain is also of great help to the research on the route selection of individual passengers in the rail transit network.

(3) Joint analysis of multicity data to improve the universality and robustness of the model: this paper takes Beijing as an example to check and verify the parameters of each model. From the results, it can be seen that the model framework has ideal prediction accuracy. However, the applicability of the model parameters to other cities and the robustness of the model's prediction accuracy to other cities cannot be estimated. Therefore, in order to improve the universality and robustness of the model and make it more suitable for engineering practice, later studies can use data from multiple cities to conduct spatial and horizontal joint analysis and verify the model parameters.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partly supported by the National Nature Science Foundation of China–Young Scientists Fund (Grant no. 71103014), the Beijing Municipal Philosophy Office (Grant no. 14JGC095), the Science and Technology Project

of the Beijing Traffic Commission (Grant no. B17M00080), and the Science and Technology Project of Beijing Transportation Industry (Grant no. 201905-ZHJC2).

References

[1] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[2] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, no. 1, pp. 132–141, 1998.

[3] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, no. 1, pp. 179–188, 1999.

[4] C. Cai, E. Yao, and M. Wang, "Passenger flow prediction of inbound and outbound stations of urban rail transit based on product ARIMA model," *Journal of Beijing Jiaotong University*, vol. 38, no. 2, pp. 135–140, 2014.

[5] M. Milenković, L. Švadlenka, and V. Melichar, "SARIMA modelling approach for railway passenger flow forecasting," *Transport*, vol. 33, no. 5, pp. 1–8, 2015.

[6] Y. Wang, B. Han, and Q. Zhang, "SARIMA model-based passenger flow prediction of Beijing subway station," *Transportation System Engineering and Information*, vol. 15, no. 6, pp. 205–211, 2015.

[7] D. Q. Wu, M. Dong, H. Y. Li, and F. Li, "Vehicle routing problem with time windows using multi-objective co-evolutionary approach," *International Journal of Simulation Modelling*, vol. 15, no. 4, pp. 742–753, 2016.

[8] J. M. Munoz-Guijosa, E. Riesco, and M. Olmedo, "Neural network and training strategy design for train drivers' vibration dose simulation," *International Journal of Simulation Modelling*, vol. 16, no. 1, pp. 72–83, 2017.

[9] J. Guo, W. Huang, and B. M. Williams, "Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[10] H. Deng, X. Zhu, and Q. Zhang, "Short-term bus passenger flow prediction based on multi-core least-squares support vector machine," *Journal of Transportation Engineering and Information*, vol. 10, no. 2, pp. 84–88, 2012.

[11] D. Y. Zhang and H. N. Yang, "Passenger flow analysis in subway using a kind of neural network," *Applied Mechanics and Materials*, vol. 715, pp. 2284–2287, 2015.

[12] S. Wang, R. Zhou, and L. Zhao, "Forecasting Beijing transportation hub areas's pedestrian flow using modular neural network," *Discrete Dynamics in Nature and Society*, vol. 2015, 6, pp. 1–6, 2015.

[13] B. Tadic, M. Zivkovic, and G. Simunovic, "The influence of vacuum level on the friction force acting on the pneumatic cylinder sealing ring," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 4, pp. 970–976, 2019.

[14] W. F. Yu, G. S. Hou, and P. C. Xia, "Supply chain joint inventory management and cost optimization based on ant colony algorithm and fuzzy model," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1729–1737, 2019.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] Y. Bai, Z. Sun, B. Zeng, J. Deng, and C. Li, "A multi-pattern deep fusion model for short-term bus passenger flow forecasting," *Applied Soft Computing*, vol. 58, pp. 669–680, 2017.

[17] G. D. Thomas, "Machine learning research: four current directions," *Ai Magazine*, vol. 18, no. 4, pp. 97–136, 1997.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[22] A. Ledezma, R. Aler, A. Sanchis, and D. Borrajo, "GA-stacking: evolutionary stacked generalization," *Intelligent Data Analysis*, vol. 14, no. 1, pp. 89–119, 2010.

[23] Y. Son and G.-G. Jin, "A nonlinear PD controller design and its application to MOV actuators," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 5–12, 2019.

[24] G. I. Y. Mustafa, H. Wang, and Y. Tian, "Model-free adaptive fuzzy logic control for a half-car active suspension system," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 13–24, 2019.

Hindawi
Complexity
Volume 2020, Article ID 6641592, 17 pages
https://doi.org/10.1155/2020/6641592

Research Article

Fund Network Centrality, Hard-to-Value Portfolio, and Investment Performance

Xiao Hu,[1] **Yimeng Cang**,[1] **Long Ren**,[2] **and Jun Liu**[3]

[1]*School of Finance, Southwestern University of Finance and Economics, Chengdu 611130, China*
[2]*School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China*
[3]*International Business School, Beijing Foreign Studies University, Beijing 100089, China*

Correspondence should be addressed to Jun Liu; liuj@bfsu.edu.cn

Received 14 October 2020; Revised 20 November 2020; Accepted 30 November 2020; Published 17 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

Based on the quarterly data of mutual funds in China from the fourth quarter of 2004 to the fourth quarter of 2019, this paper constructs a series of complex bipartite networks based on the overlapped portfolios of mutual funds and then explores the influences of fund network position on mutual fund's investment behavior and performance. This paper finds that a mutual fund with shorter information transmission path to other entities in the fund network (i.e., having higher closeness centrality) or with stronger ties with those entities in important information positions (i.e., having higher eigenvector centrality) will achieve better investment performance. However, a stronger mediating role over the potential information flow of the fund network (i.e., having higher betweenness centrality) cannot help a mutual fund increase performance. The empirical results also indicate that a mutual fund holding stock portfolios with high valuation difficulties caused by the market or fundamental information uncertainty will achieve better investment performance, while holding hard-to-value portfolios caused by limited public information will reduce the performance of the fund. Furthermore, high closeness centrality or eigenvector centrality can help mutual funds deal with the disclose problems of public information, thus reducing the likelihood of a mutual fund holding hard-to-value portfolios caused by limited public information to achieve worse performance. Eigenvector centrality brings information advantages about company fundamentals, so it is easier for a mutual fund with high eigenvector centrality to profit from holding hard-to-value portfolios caused by the fundamental information uncertainty. The conclusions of this paper can enhance our understanding of the fund network and its information mechanism and shed new light on mutual fund's information advantages and related asset allocation strategies.

1. Introduction

Investors need to rely on information to make investment decisions. Since mutual funds are main institutional investors in the capital market, how they obtain and use information advantages have attracted scholars' attention. Many studies indicate that information advantages help mutual funds improve investment performance [1–4].

One important way for a mutual fund to gain information advantages is through the fund network. The network reflects a collection of social interaction relations and provides a channel for information transmission among social actors [5–10]. A mutual fund is embedded in a certain network where it directly or indirectly associates with other entities [11–13]. Therefore, it can obtain valuable information related to investment by interacting with others in the fund network, thus shaping their investment strategies and returns [14, 15]. In other words, the fund network exerts influences on mutual funds through the information mechanism of the network [16].

A network structure consists of nodes and edges, in which the nodes are actors in social interactions and the edges are the connections between nodes through such social activities [5]. From the perspective of network embeddedness, each node in the network plays the two roles of the mediator and filter, for network information flows [17]. Thus, the location of a node in the network determines the quantity and quality of information; it can access from the network [9]

as well as has the ability to address information [18]. Accordingly, a better-networked mutual fund with a more influential position in the fund network may occupy more information advantages, which can improve its investment performance. Graph theory provides us with tools for measuring the importance of the network position of each node relative to other nodes. The tools mainly involve four network indicators: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. These four kinds of network centrality have different definitions. Degree centrality is the simplest indicator and counts the number of direct ties of a focal network node [19]. The nodes directly connected with more nodes have more channels to obtain information and obviously have more favorable network positions. Closeness centrality is the reciprocal of the average shortest path of a focal node's access to other nodes directly or indirectly in the network [20]. It shows the connectivity of a node in the network and can reflect the efficiency of the node to obtain information by using its network location. Betweenness centrality measures the possibility that a focal node lies on the shortest path of any two other nodes, so it can indicate the mediating role over the potential information flows in the network that the focal node plays [21]. Eigenvector centrality is calculated from the eigenvectors of the network adjacency matrix and represents the degree to which a focal node builds ties with influential nodes in the network [22]. Most literature studies on the position of mutual fund in the fund network focus on degree centrality [12, 23] and support that degree centrality could improve fund performance [24]. However, degree centrality neither considers the indirect ties in the fund network nor can capture the network structural characteristics contained in the other three centrality indicators. To this end, this paper explores the influences of closeness centrality, betweenness centrality, and eigenvector centrality of fund network on fund performance, especially compares the differences in the effects of the three network centrality indicators.

Although previous studies analyze the relationship between fund network and fun performance based on the information functions of the network [11, 25], the underlying mechanism by how the information advantages brought by the network position help a mutual fund improve investment performance has not been examined. This paper tries to deal with the research gap by testing the impacts of a mutual fund's network centrality on the performance of its investment behaviors. Specifically, we focus on the investment behavior in which mutual funds allocate stocks with high valuation difficulty. High information uncertainty, poor disclosure, or low quality of public information will increase the difficulty of stock valuation, causing it hard for mutual funds to make accurate predictions on such stocks [26–28]. Holding hard-to-value stock portfolios require mutual funds to have stronger capabilities of information mining and analysis to reduce investment risks. Moreover, hard-to-value stocks are more likely to be mispriced, so they always have richer arbitrage opportunities [29–37]. The information advantages possessed by better-networked mutual funds may help generate greater returns from the hard-to-value portfolios.

Using the quarterly data of China's mutual funds from the fourth quarter of 2004 to the fourth quarter of 2019, this paper explores three important questions: (i) how closeness centrality, betweenness centrality, and eigenvector centrality in fund network impact mutual fund's investment performance, respectively; (ii) how holding hard-to-value portfolio impacts mutual fund's investment performance; (iii) how the information advantages brought by the three network centrality indicators impact mutual fund's return from the hard-to-value portfolio.

2. Hypothesis Development

2.1. Fund Network Centrality and Investment Performance. Network centrality reflects the importance of the network position of each node relative to other nodes. We argue that a mutual fund with higher centrality in the fund network can achieve better investment performance. There are two main reasons. First, a mutual fund centrally located in the fund network can connect with other entities through fewer intermediaries; this can improve their efficiency and reduce their costs to acquire information from the fund network and help them overcome the information constraints that other entities impose [38]. In addition, the more influential the position of a mutual fund occupies in its network, the higher the level of involvement it has in the information flow through direct or indirect ties with others [39, 40]. This will help the mutual fund access more information sources [41, 42] while also ensuring the quality of information by comparing different sources [18].

It also should be noted that closeness centrality, betweenness centrality, and eigenvector centrality captures different structural features of the fund network, so they may have different influences on mutual funds. Closeness centrality reflects the length of the information transmission path between a mutual fund and other entities in the fund network [43]. Betweenness centrality reflects the mediating effect of a mutual fund on the potential information flow in the fund network [44]. Eigenvector centrality in the fund network reflects the degree of relationships between a mutual fund and other entities in important information positions [45]. Therefore, we believe that the three types of network centrality bring different information advantages to mutual funds. As such, Hypothesis 1 is as follows.

Hypothesis 1. A mutual fund with a more central position in the fund network will achieve better investment performance, but the effects of closeness centrality, betweenness centrality, and eigenvector centrality could be different.

2.2. Hard-to-Value Portfolio and Investment Performance. Stocks with higher valuation difficulty have less public information, less transparency, and more information uncertainty, and meanwhile, their related indicators such as company fundamentals and market sentiment have higher volatility. Because hard-to-value stocks are more likely to be mispriced, mutual funds may use the potential arbitrage opportunities to achieve better performance. According to

Lai et al. [36] and Kumar [33], mature institutional investors have advantages in terms of analytical capability and information acquisition, so they are more likely to make profits by taking advantage of pricing errors when information is uncertain. Chen et al. [46] study the trading situations of active funds and point out that the trading behavior of a fund reflects the fund's ability to identify mispriced stocks. Therefore, we argue that mutual funds are more likely to profit from their active allocation on hard-to-value portfolios.

There are three main reasons for the valuation difficulty of stocks, generating three types of hard-to-value portfolios. The first is caused by market information uncertainty. Although the trading price of a stock can be observed by investors, the high volatility and low synchronicity of its price make it difficult to be evaluated [33]. The second is caused by fundamental information uncertainty. Good financial condition and corporate governance of a listed company are helpful for mutual funds to accurately analyze its fundamentals with public information [27]. The third is caused by limited public information. The disclosure of a listed company's public information such as annual reports and research reports provided by third-party institutions also affects its valuation difficulties [47]. The information attributes behind the three types of hard-to-value portfolios are different. In such cases, we propose Hypothesis 2 as follows.

Hypothesis 2. Holding hard-to-value portfolios can help a mutual fund achieve better investment performance, but there could be different returns of the hard-to-value portfolios caused by market information uncertainty, fundamental information uncertainty, or limited public information.

2.3. Fund Network Centrality and Return of Hard-to-Value Portfolio. Holding hard-to-value portfolios require mutual funds to have stronger capabilities of information mining and analysis to reduce investment risks and take full use of the potential arbitrage opportunities. Mutual funds with strong information advantage are abler to make profits from the portfolios with high valuation difficulty. Since closeness centrality, betweenness centrality and eigenvector centrality bring some information advantages to a mutual fund, they will affect the mutual fund's information mining and information analysis [3, 4], thereby influencing the returns of hard-to-value portfolios. Further considering the different information advantages brought by the three-fund network centrality indicators and the different information attributes of the three types of hard-to-value portfolios, Hypothesis 3 is proposed as follows:

Hypothesis 3. Fund network centrality increases the likelihood for a mutual fund to achieve better investment performance by holding hard-to-value portfolios, but the three centrality indicators could have different influences on the returns of the three types of hard-to-value portfolios.

3. The Data

This paper uses open-ended stock funds or stock-leaning funds which execute active strategies in China as the research object. Although our empirical data is about Chinese mutual funds, we do not want to highlight the Chinese research context too much but try to make general contributions to existing studies. Admittedly, as an emerging market, China's stock market has higher information uncertainty [48], more mispricing phenomena [49], and speculative investment behaviors than mature markets [50]. It may be more effective for Chinese mutual funds to use a fund network and invest hard-to-value stocks to improve investment performance. However, previous studies also indicate that even the mature capital markets of developed countries cannot reach the standard of weak efficiency, so stronger capabilities of information mining and analysis can help fund managers to achieve better performance [51]. Besides, a fund network has been proven to exist in mature markets and can be an important way for mutual funds to obtain information advantages [24]. To this end, the Chinese market is just an appropriate research context where information is more valuable for investment, and the conclusions are drawn from the market and still have applicability to other markets.

Since China's open-ended fund market was formed at the end of 2004, the sample period is from the fourth quarter of 2004 to the fourth quarter of 2019. We exclude funds that are Shanghai-Hong Kong Stock Connect funds and Qualified Domestic Institutional Investor (QDII) from our samples because these funds may invest overseas stocks which could bring difficulties for us to construct fund networks and compare investment behaviors and performance among funds. To eliminate the possible interferences of mutual funds' initial investment period of building up positions, mutual funds established less than one year are further excluded. When constructing quarterly fund networks, several mutual funds cannot access any other funds in the networks because their portfolios have no overlaps with others. Calculating the centrality of unconnected networks may distort the indicators of corresponding nodes, so we only retain the mutual funds in the largest connected networks. The final sample contains 2,802 funds and 37,607 fund-quarter observations.

The data of each mutual fund's quarterly portfolios and other details about the fund manager, fund financial statement, and trading information are obtained from the mutual fund's first quarter, semiannual, third quarter, and annual reports collected by the CSMAR database. We obtain most of the stock data such as ROE ROA, operating income, market value, and daily stock price from the RESSET database. Other data used in the paper such as audit data and securities analyst data of each stock are from the CNRDS database.

4. Fund Network Construction

Previous studies employ three methods to construct the fund network. The first is to define network ties based on the same

education or career background of fund managers. For example, Shen [25] indicates that the alumni network enables fund managers to obtain private information, thus promoting the performance of the fund. The second is to define network ties based on the geographical location of fund managers. Hong et al. [11] find that geographic proximity would make it possible for individuals to exchange more information, so the trading behavior of a fund manager would be significantly affected by other fund managers in the same city. The above two types of fund network are intuitive, but the information transmission in the network cannot be observed directly because the ties between mutual funds are defined from communication sources. The relations based on geographical location or background do not mean that fund managers actually have access to information exchanges that can influence their investment decisions.

In this regard, some scholars construct the fund network from the overlapped portfolios of mutual funds. According to Bushee and Goodman [14], Jiang [15], and Cohen et al. [52], when a mutual fund holds a certain stock in a large position, it means that the mutual fund may own more information of that stock. Shiller and Pound [53] believe that a mutual fund may have some information exchange with other mutual funds which hold the same stocks, and its investment decisions are influenced by them. Pareek [12] indicates that mutual funds who hold the same stocks show significantly consistent investment behavior, which could not be explained by the funds' investment style and geographical location. To some extent, the fund network based on the overlapped portfolios of mutual funds is not limited to a certain type of information connection because it covers the possible information exchanges brought by geographical proximity or the same background. In other words, if the information advantages obtained from others who have a near geographical location or the same background can influence a fund manager's investment decision, it will eventually be reflected in the portfolio behavior of the fund manager. In addition, the fund shareholding network shows the relations between mutual funds established by their overlapped portfolios, which is more directly related to mutual funds' investment behaviors and performance. Compared with the other two kinds of fund networks, the fund shareholding network may be a more significant driven factor for fund behavior and performance.

Expect for the social relations between mutual funds, mutual funds and stocks they hold may interact with each other [54, 55]. Cohen et al. [56] indicate that fund managers are more willing to invest in stocks whose executives have the same education experience. Frankel et al. [57] find that, during the teleconference, there are increasing volume and price volatility of the listed company's stock because institutional investors may use the message of the teleconference for trading. Since mutual funds can both obtain information about stock from other mutual funds and listed companies, we use a bipartite network approach to comprehensively describe the information channels of mutual funds [58].

In bipartite networks, actors of a common nature are linked through their joint actions known as events, and the events themselves exist in the network as nodes. When two mutual funds invest in the same stock, the stock acts as an event node that can link the two funds. The bipartite network method helps us capture the interdependencies between the mutual funds and the stocks [59] and can reflect the information advantages of mutual funds obtained from the two sources of listed companies and mutual funds. Specifically, we define two types of nodes: mutual funds and stocks. At the end of each quarter, if the market value of a stock held by a mutual fund accounts for more than 5% of its position1, a tie is established between them. When we take each stock as a subset of a group of funds, a bipartite network at that time is constructed.

Figure 1 shows the bipartite network on 31 March 2005. The blue nodes represent mutual funds, and the white nodes represent stocks. The edges between a blue node and a white node indicate that a fund holds stock in more than 5% of its position. Figure 2 is a part of the network in Figure 1 and shows all direct connections of fund 040002 (the red node). The mutual fund held five stocks (i.e., 000792, 000402, 000538, 000866, and 600096) in a large position. The five stocks were the main holdings of the other 25 funds. In other words, there were connections for Fund 04002 to the 25 funds through the five stocks in its portfolio.

5. Measurement

5.1. Investment Performance. Following the method of Agarwal et al. [60], we use Alpha after risk adjustment for the four-factor model to measure the investment performance of mutual funds [61]. In the first step, the monthly return rate of the previous 24 months of mutual fund j is used for regression to calculate the four-factor parameters. The regression equation is as follows:

$$R_{i,s} = \widehat{\alpha}_{i,t-1} + \sum_{k=1}^{4} \widehat{\beta}_{i,k,t-1} F_{k,s} + \varepsilon_{i,s}, \quad s = m-24, \ldots, m-1, \tag{1}$$

where s and m stand for month, R stands for fund i's the monthly return on net value excluding the risk-free interest rate, And F stands for monthly four factors (i.e., market risk premium factor, market value factor, book value factor, and momentum factor).

The monthly excess return rate of the fund is calculated by the following formula:

$$Performance_{i,m} = R_{i,m} - \sum_{k=1}^{4} \widehat{\beta}_{i,k,m-1} F_{k,m}. \tag{2}$$

5.2. Closeness Centrality. In a one-mode network, closeness centrality is the reciprocal of the average shortest path of focal node access to other nodes in the network. According to Faust [62], the actor nodes of a bipartite network are only adjacent to event nodes, and all paths emanating from an actor must first pass through the events to which the actor belongs. To this end, the closeness centrality of mutual fund i

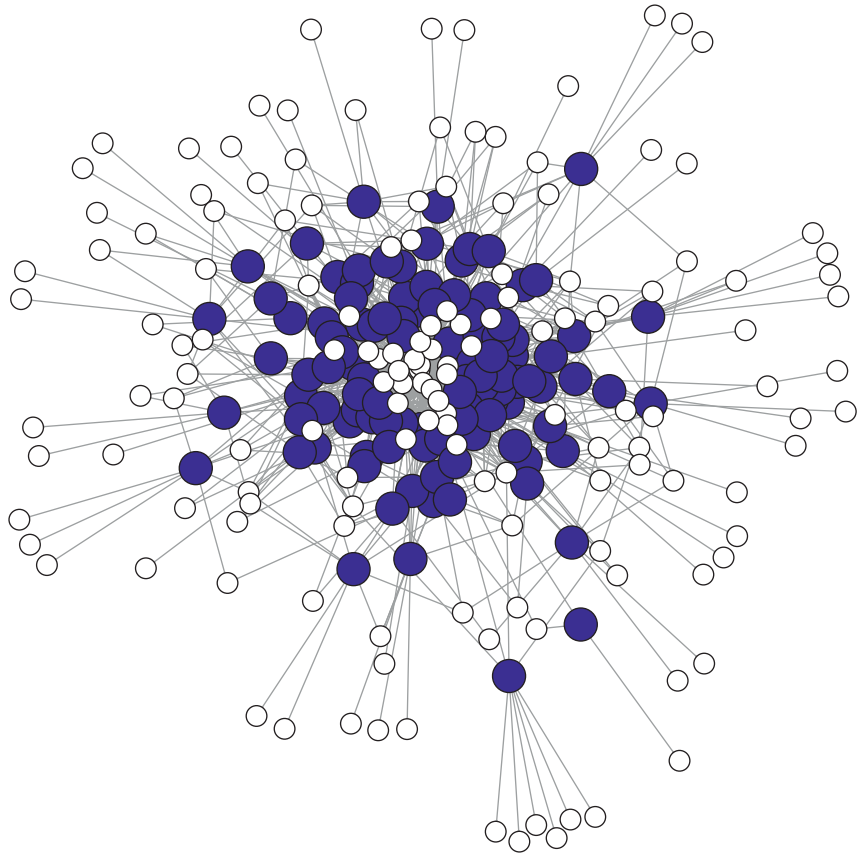


Figure 1: The bipartite network on 31 March 2005.

in our bipartite network is a function of the minimum distances from any of its stocks and to other funds and stocks in the network:

$$Closenss_i = \left[1 + \frac{\sum_{j=1}^{g+h} \min_k d(k, j)}{g + h - 1}\right]^{-1}, \quad (3)$$

where k are the stocks connected to fund i, $d(k, j)$ represents the minimum distance between stocks k and other funds or stocks in the network, and g and h, respectively, represent the number of funds and stocks.

5.3. Betweenness Centrality. Betweenness centrality for a one-mode dyadic network focuses on the extent to which nodes sit on geodesic paths between other pairs of nodes [21]. In a bipartite network, linkages between pairs of event nodes are always through the joint memberships of actor nodes, thus actor nodes are always on paths between events. In calculating the betweenness centrality of mutual fund i, in a bipartite network, we focus on the collection of stocks that belong to the fund. Fund i is on a path between all pairs of actors that are members of it. If a given pair of stocks, (l, k), only shares fund i in common (thus $x_{kl}^M = 1$), then fund i is on the only path between them, and fund i's betweenness centrality is incremented by $(1/x_{kl}^M)$ for each pair of stocks (l, k) in fund i. Thus, a portion of the betweenness centrality of fund i can be expressed as follows:

$$Betweenness_i = \frac{1}{2} \sum_{m_k, m_l \in n_i} \frac{1}{x_{kl}^M}. \quad (4)$$

In addition, an event gains betweenness centrality if a stock belongs only to fund i. In that case, all paths from the stock must contain the fund. Since there are $g + h$ nodes (funds and stocks) in the bipartite network, a fund gains $g + h - 2$ betweenness centrality points' for each of its members that belong to no other funds. This quantity is not independent of the count in equation (4).

5.4. Eigenvector Centrality. Eigenvector centrality provides a measure of the extent to which a focal node builds ties with influential nodes in the network [22]. Bonacich [63] shows that, for a bipartite network, the eigenvector centrality of actor nodes and event nodes can be derived from the following equation:

$$\lambda \begin{bmatrix} \mathbf{c}^N \\ \mathbf{c}^M \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{A} \\ A' & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c}^N \\ \mathbf{c}^M \end{bmatrix}, \quad (5)$$

where $\boldsymbol{A} = \{a_{ij}\}$ ($i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, h$) is denoted as the adjacency matrix between actor nodes and event nodes. g and h are the number of actor nodes and event nodes, respectively. λ is the largest eigenvalue of matrix $\mathbf{A}.\mathbf{c}^N$ and $\mathbf{c}^M$ are the eigenvectors corresponding to the largest eigenvalue for actor nodes and event nodes, respectively.

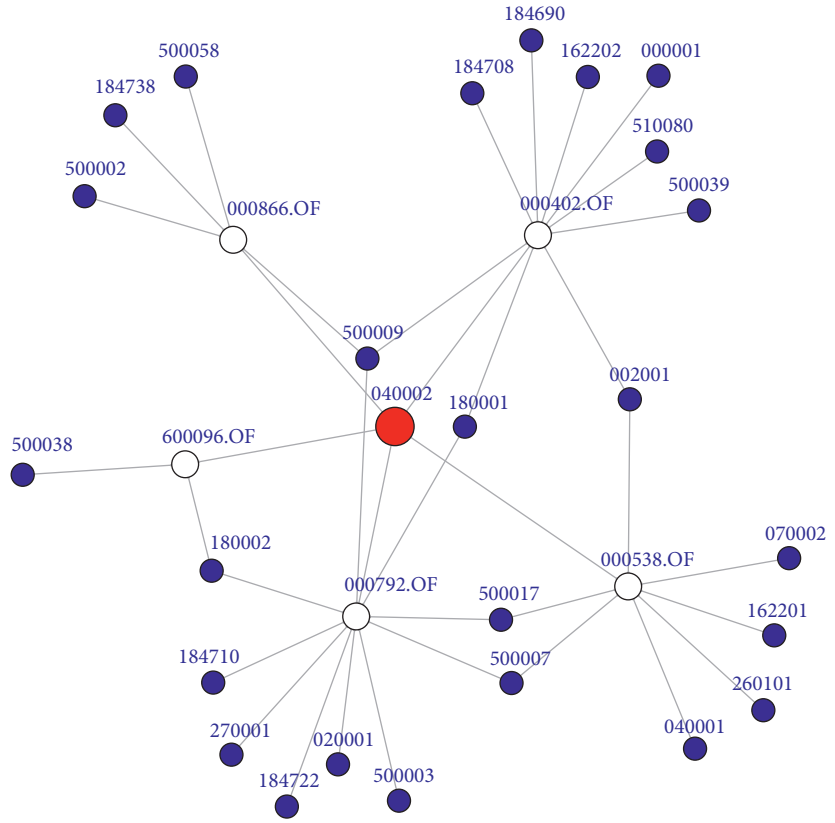


Figure 2: The connections of fund 040002 (the red node) on 31 March 2005.

When we denote **A** as the adjacency matrix between mutual funds and stocks, the eigenvector centrality of mutual fund i ($Eigenvector_i$) can be obtained from the vector of eigenvector centrality scores for funds $\mathbf{c}^N$.

5.5. Hard-to-Value Portfolio Caused by Market Information Uncertainty. We use four indicators to measure the hard-to-value portfolio caused by market information uncertainty. The first indicator is stock price volatility, which is obtained by dividing the standard deviation of a stock's weekly return rate in the half-year formative period by its initial stock price. The second indicator is the stock characteristic volatility. A Fama–French three-factor model is used to carry out regression on the weekly return series of the formative period, and the standard deviation is calculated from the residual. The two indicators reflect information uncertainty. High stock price volatility and characteristic volatility may increase the valuation difficulty of a stock.

The third indicator is the R^2 of the regression results using asset pricing model. The following equation shows the regression with the data of the weekly return rate:

$$r_{i,t} = \alpha_i + \beta_i r_{m,t} + c_i I_{i,t} + e_{i,t}, \tag{6}$$

where $r_{i,t}$ is the return rate of stock i in week t, $r_{m,t}$ stands for the return rate of market in week t and $I_{i,t}$ stands for the return rate of industry in which stock i belongs. The R^2 value after regression measures the part of stock i's return that can be explained by systematic pricing factors. The smaller R^2 reflects higher proportion of the part that cannot be predicted by systematic risks, thus increasing the valuation difficulty.

The fourth indicator is illiquidity index. The Amihud illiquidity ratio [64] is significantly and positively correlated with the information uncertainty of a stock [65]. The formula used to calculate the Amihud illiquidity of stock i in period t is as follows:

$$Illiquidity_{i,t} = \frac{1}{n_{i,t}} \sum_{d=1}^{n_{i,t}} \frac{|R_{i,t,d}|}{V_{i,t,d}}, \tag{7}$$

where $n_{i,t}$ represents the trading days of stock i in period t, $R_{i,t,d}$ represents the return rate of stock i on day d, and $V_{i,t,d}$ stands for the trading volume of stock i on day d (in units of billion yuan). High illiquidity means high degree of information asymmetry, which makes a stock hard to be analyzed.

We sort all stocks according to the ascending order of stock price volatility, stock characteristic volatility, and illiquidity index and the descending order of R^2. The ranking of each stock obtained after sorting is divided by the number of stocks to obtain a stock's hard-to-value score at the four dimensions. By adding up the four hard-to-value scores, we can get the score of stock. Taking the proportion of a stock's market value in a mutual fund's investment portfolio as the weight, the valuation difficulty of the fund's portfolio caused by market information uncertainty (HTV_mrkt) is the weighted average of all stocks' scores it holds.

5.6. Hard-to-Value Portfolio Caused by Fundamental Information Uncertainty. We use seven indicators to measure the hard-to-value portfolio caused by fundamental information uncertainty. The first indicator is the stock size. Small-sized stocks have high information asymmetry and are easier to be affected by market sentiment, increasing their valuation difficulty [66]. We measure stock size as the market value. The second indicator is the proportion of fixed assets to the total asset. According to Baker [34], the tangibility of a company's assets will affect the difficulty of its valuation. The higher the proportion of intangible assets in a company, the more difficult it is to measure its value. Besides, if a company spends more on research and development activities, it is harder to estimate the expected consequences, causing more valuation uncertainty. To this end, the proportion of fixed assets to the total asset has a negative relationship with a stock's valuation difficulty. The third indicator is the volatility of operating income, which is measured as the standard deviation of a stock's operating income in the previous eight quarters. High volatility of operating income increases the valuation difficulty of a stock.

The fourth indicator is the book-to-market ratio. According to Baker [34] and Kumar [33], a lower book-to-market ratio means higher potential growth of a stock and may increase the valuation difficulty. The fifth and sixth indicators are ROA volatility and ROE volatility, which are measured as the standard deviation of a stock's ROA and ROE for the previous eight quarters, respectively. High volatility of ROA and ROE means low profitability stability, bringing difficulties in valuing a stock.

The seventh and eighth indicators are the share proportion of the largest shareholder and the share proportion of the second to the tenth largest shareholder. La Porta et al. [67] indicate that over dispersed equity is likely to cause serious agency problems and may increase the degree of information asymmetry of a company, thus increasing its difficulty of valuation. The lower the share proportion of the largest shareholder is and the higher the share proportion of the second to the tenth largest shareholder, the more difficult it is to assess a stock's value.

We sort all stocks according to the descending order of stock size, proportion of fixed assets to total asset, book-to-market ratio, and share proportion of the largest shareholder and the ascending order of the volatility of operating income, ROE, ROA and share proportion of the second to the tenth largest shareholder. The ranking of each indicator obtained after sorting is summed and divided by the number of stocks to obtain a stock's hard-to-value score. We measure the valuation difficulty of a mutual fund's portfolio caused by fundamental information uncertainty (*HTV_fdmt*) as the weighted average of all stocks' hard-to-value scores it holds.

5.7. Hard-to-Value Portfolio Caused by Limited Public Information. Two indicators are used to measure the hard-to-value portfolio caused by limited public information. The first indicator is the number of analysts who follow a stock in the previous year. Lang and Lundholm [68], Hong et al. [69], and Gleason and Lee [70] all believed that more analysts following a stock will increase its information disclosure and reduce the information asymmetry, which is helpful for reducing the valuation difficulty of the stock.

The second indicator is the size of a stock's accounting firm. Audit reports issued by larger accounting firms usually have higher quality, which can reduce the difficulty of stock valuation. We measure the size of an accounting firm as its annual revenue.

All stocks are sorted according to the descending order of the number of analysts and the size of accounting firm. We then use the methods mentioned above to calculate the hard-to-value score of a mutual fund's hard-to-value score at the public information disclosure aspect (*HTV_info*).

5.8. Control Variables. We include two sets of control variables related to fund manager and mutual fund that might affect the performance of mutual funds. According to Simutin [71] and Niessen and Ruenzi [72], fund manager's education (*PhD*) and gender (*Gender*), tenure length (*Tenure*), and fund management team size (*Teamsize*) are controlled at the fund manager aspect. *PhD* is a dummy variable. If at least one fund manager of a mutual fund holds a PhD in the current fund management team, the value is 1; otherwise, it is 0. *Gender* is also a dummy variable. If at least one fund manager in the current fund management team is female, *Gender* equals to 1; otherwise, it equals to 0. We measure *Tenure* as the natural logarithm of the average tenure (in units of month) of all managers who manage a mutual fund in the current period. *Teamsize* is measured as the number of fund managers who jointly manage a mutual fund in the current period.

Following the literature [73–75], we control for factors related to mutual fund characteristics, including fund size (*Netasset*), fund family size (*Familysize*), fund age (*Fundage*), fund's expense ratio (*Expense*), turnover ratio (*Turnover*), historical performance (*Return*), fund flow (*Fundflow*), and investment style fixed effect (*Style*). *Netasset* is equal to the natural logarithm of the total net assets of a mutual fund at the end of the previous period. We measure *Familysize* as the natural logarithm of the total net assets of the fund family to which a mutual fund belongs at the end of the previous period. *Fundage* is equal to the natural logarithm of the interval (in units of month) from the founding date of a mutual fund to the end of the previous period. The ratio of the total operating expenses of a mutual fund for the previous period to the average total net assets of the fund at the beginning and ending of the previous period is adopted to measure *Expense*. *Return* is measured as the return on equity of the last period of a mutual fund. We use the net capital inflow (in units of billion yuan) in the previous period of a mutual fund to measure *Fundflow*.

We include a series of dummy variables to control the investment style fixed effect of a mutual fund. Referring to Daniel et al. [76], Pareek [12], and Hoberg et al. [77], the paper describes the investment style of a mutual fund through three dimensions of the stocks in its portfolio: size, book-to-market ratio, and momentum. The current market value (in units of billion yuan) is used to represent the size of

a stock, and the cumulative return rate in the previous year (excluding the last month to avoid the effect of short-term reversal of price) is used to represent the momentum of a stock. In order to avoid the interference of outliers, the above three indicators are all winsorized by 1%. There are four steps to further construct the dummy variables of investment style. First, we take the natural logarithm of the three dimensions of a stock and then standardize them to get the transformations of size ($z\ln Size$), book-to-market ratio ($z\ln BM$), and momentum ($z\ln Mom$). Second, cross-sectional regressions are made for all stocks in each period to obtain a series of regression coefficients:

$$\begin{aligned} z\ln BM_i &= \alpha + \beta z\ln Size_i + \varepsilon, \\ z\ln Mom_i &= \delta + \theta z\ln Size_i + \gamma z\ln BM_i + \varepsilon. \end{aligned} \tag{8}$$

The book-to-market ratio and momentum after orthogonalization adjustment are

$$\begin{aligned} rz\ln BM_i &= z\ln BM_i - \alpha - \beta z\ln Size_i, \\ rz\ln Mom_i &= z\ln Mom_i - \delta - \theta z\ln Size_i - \gamma z\ln BM_i. \end{aligned} \tag{9}$$

Third, taking the proportion of a stock's market value in a mutual fund's investment portfolio as the weight, the three indicators of the fund's investment style (i.e., *fSize*, *fBM*, and *fMom*) are the weighted average of all stocks' $\ln Size$, $rz\ln BM$, and $rz\ln Mom$. Fourth, all funds are assigned with the three style scores. Each style score is a scale of 1 to 5 to indicate the ranking of the corresponding indicator in the top 20%, lower than 20% but higher than 40%, lower than 40% but higher than 60%, lower than 60% but higher than 80%, or last 20%. The combinations of the three style scores generate 125 types of investment styles, and 125 style dummy variables are defined. For example, if a fund ranks 20–40% in *fSize*, top 20% in *fBM*, and last 20% in *fMom*, its investment style can be described as "2-1-5" and the corresponding dummy variable equals to 1.

A series of observation quarter dummies are also used to control for time fixed effect (*Quarter*), which allows us to reduce the unobservable effects of time.

5.9. Statistical Summaries. Table 1 reports the descriptive statistics and correlation matrix for the key variables. For clarity and the convenience of interpretation, we use the raw values of *Netasset, Familysize, Fundage*, and Tenure instead of their natural log transformations in Table 1. On average, a mutual fund has 1.73 billion net assets, belongs to a fund family with 44.10 billion net assets, and obtains a 0.44% adjusted monthly excess return rate. The average fund age is 57.30 months and the average tenure of the fund manager is 26.39 months. 15.9% of mutual funds have managers with Ph.D. and 22.33% have a female manager. On average, a mutual fund has 1.39 fund managers, a 2.69% expense ratio, and a 195.7% turnover rate.

According to the correlation matrix, apart from the significant correlation between *Closeness*, *Betweenness*, and *Eigenvector*, there is no obvious collinearity between variables. In fact, we do not put the three centrality indicators in one model. The relatively high correlation between the three network centrality is not difficult to understand. The three centrality indicators measure the importance of a node relative to other nodes from different aspects, and the different information advantages driven by the different types of influential positions may convert mutually to some extent [62, 78]. For example, a node with good relations with nodes having important information positions (i.e., high eigenvector centrality), may construct highly efficient channels to access more information sources (i.e., high closeness centrality) with the help of the influential nodes, and may become information broker of the network information flow (i.e., high betweenness centrality). However, this paper focuses on the different natures of the three centrality indicators, and the following empirical results indicate their differences even if there are high correlations between them.

6. Fund Network Centrality and Investment Performance

6.1. Model Specification. To test the relationship between fund network centrality and investment performance, we employ the following fixed effect model and adjust the t-value by clustering the observations by fund code:

$$\begin{aligned} Performance_{j,t} = {} & \alpha + \beta_1 \cdot Closeness_{j,t-1} + \beta_2 \cdot Netasset_{j,t-1} \\ & + \beta_3 \cdot Familysize_{j,t-1} + \beta_4 \cdot Fundage_{j,t-1} \\ & + \beta_5 \cdot Expense_{j,t-1} + \beta_6 \cdot Turnover_{j,t-1} \\ & + \beta_7 \cdot \mathrm{Return}_{j,t-1} + \beta_8 \cdot Fundflow_{j,t-1} \\ & + \beta_9 \cdot PhD_{j,t} + \beta_{10} \cdot Gender_{j,t} \\ & + \beta_{11} \cdot Teamsize_{j,t} + \beta_{12} \cdot Tenure_{j,t} \\ & + \sum_i \lambda_i Style_{i,j,t} + \sum_k \gamma_k Quarter_{k,t} + \varepsilon_{j,t}, \end{aligned} \tag{10}$$

where $Performance_{j,t}$ represents the investment performance of fund j in quarter t, $Closeness_{j,t-1}$ represents the closeness centrality of fund j in quarter $t-1$, which can be replaced by $Betweenness_{j,t-1}$ or $Eigenvector_{j,t-1}$ when the effects of betweenness centrality or eigenvector centrality is tested, $Netasset_{j,t-1}$, $Familysize_{j,t-1}$, $Fundage_{j,t-1}$, $Expense_{j,t-1}$, $Turnover_{j,t-1}$, $Return_{j,t-1}$, and $Fundflow_{j,t-1}$ refer to fund j's net asset, total net asset of its family fund, fund age, expense ratio, turnover ratio, historical performance, and net capital inflow in quarter $t-1$, $PhD_{j,t}$, $Gender_{j,t}$, $Teamsize_{j,t}$, and $Tenure_{j,t}$ refer to fund managers' education, gender, number of people, and tenure of fund j in quarter t, $Style_{j,t}$ refers to the investment style fixed effect of fund j in quarter t, $Quarter_t$ refers to the time fixed effect of quarter t, and $\varepsilon_{i,t}$ denotes the error term.

6.2. Results. The regression results are shown in Table 2. Model (1-1) shows that the effect of *Closeness* on *Performance* is positively significant ($\beta > 0$; $p < 0.05$), indicating that mutual funds with high closeness centrality will achieve

Table 1: Descriptive statistics and correlation matrix.

	Variable	Mean	STD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	*Performance*	0.004	0.033	−0.271	0.356																	
2	*Closeness*	0.016	0.018	0.003	0.161	0.01																
3	*Betweenness*	0.002	0.003	0	0.064	−0.02*	0.54*															
4	*Eigenvector*	0.017	0.018	0	0.151	0.01	0.56*	0.33*														
5	*HTV_mrkt*	1.095	0.602	0.006	3.079	−0.01*	−0.10*	−0.01*	−0.15*													
6	*HTV_fdmt*	2.585	1.316	0.065	5.345	−0.03*	−0.09*	−0.04*	−0.06*	0.09*												
7	*HTV_info*	0.394	0.239	0.002	1.657	−0.04*	−0.16*	−0.01*	−0.22*	0.09*	0.08*											
8	*Netasset (billion)*	1.732	2.721	0.009	15.423	0.00	0.34*	0.11*	0.28*	−0.08*	−0.05*	−0.12*										
9	*Familysize (billion)*	44.099	42.089	0.369	191.659	0.02*	−0.09*	−0.10*	0.01	−0.02*	0.01	−0.02*	0.28*									
10	*Fundage (months)*	57.306	41.092	0	216	−0.01*	−0.16*	−0.11*	−0.07*	0.01*	0.00	0.01*	0.16*	0.09*								
11	*Expense*	0.027	0.034	0.003	0.296	−0.04	−0.08	−0.02	−0.04	0.04*	0.02*	0.04*	−0.16	−0.10	−0.11							
12	*Turnover*	1.957	2.918	0.043	19.985	−0.05*	−0.11*	−0.02*	−0.10*	0.04*	0.00	0.05*	−0.22*	−0.16*	−−0.09*	0.08*						
13	*Return*	0.011	0.060	−0.163	0.187	0.01	0.02*	−0.01*	0.03*	−0.11*	−0.11*	−0.08*	0.02*	0.05*	0.02*	−0.06*	−0.03*					
14	*Fundflow (billion)*	−0.041	0.393	−1.662	1.956	0.01*	−0.03*	−0.02*	−0.04*	−0.01*	−0.01*	−0.01	−0.01*	−0.01*	−0.03*	−−0.05*	−0.00	0.05*				
15	*PhD*	0.159	0.366	0	1	−0.01	0.01*	0.02*	0.01	−0.01*	−0.01*	0.00	0.02*	−0.02*	−0.04*	0.02*	0.04*	−0.00	−0.01*			
16	*Gender*	0.223	0.417	0	1	−0.00	−0.07*	−0.05*	−0.01*	−0.01	0.02*	−0.01*	−0.09*	0.02*	−0.04*	0.02*	−0.05*	−0.01*	0.01*	−0.03*		
17	*Teamsize*	1.387	0.617	1	6	−0.01	0.01	−0.02*	0.02*	−0.00	0.01*	−0.00	0.07*	0.06*	−0.05*	0.05*	−0.03*	−0.18	−0.03*	0.14*	0.28*	
18	*Tenure (months)*	26.391	19.340	0.500	169	0.00	−0.12*	−0.10*	−0.04*	−0.01	0.02*	−0.00	0.12*	0.13*	0.34*	−0.08*	−0.10*	0.01	0.02*	−0.06*	−0.04*	−0.19*$^{\&}$

Notes: $N = 37606$; * $p < 0.1$.

better investment performance. A mutual fund with higher closeness centrality can connect with other entities through shorter information transmission paths, which can increase its efficiency in obtaining public and private information from the fund network [38]. Therefore, high closeness centrality can enhance the breadth of information acquired by a mutual fund, thus improving its investment performance.

Model (1-3) shows that the effect of *Eigenvector* on *Performance* is positively significant ($\beta > 0$; $p < 0.1$), indicating that mutual funds with high eigenvector centrality will obtain better investment performance. The higher the eigenvector centrality of a mutual fund is, the more information it can get from other entities having important information positions [79, 80]. Information from key information entities usually has high quality and high timeliness and sometimes includes potential private information. As a result, a mutual fund with higher eigenvector centrality has more information depth, which is helpful to improve investment performance.

The coefficient of *Betweenness* is not significant at the 10% confidence level in Model (1-2), indicating that betweenness centrality cannot explain for fund performance. The betweenness centrality reflects the mediating effect of a mutual fund on the potential information flow in the fund network [21], rather than the capabilities of using the network location to obtain information. Mutual funds are not financial intermediaries such as investment banks, so betweenness centrality may not give them the information advantages that can be used to improve investment performance.

In conclusion, the above results, to some extent, support Hypothesis 1 that more central position in the fund network help a mutual fund achieve better investment performance, but the effects of closeness centrality, betweenness centrality, and eigenvector centrality are different. The results also show that shorter fund age, larger fund size and fund family size, higher turnover rate, and longer tenure of the present fund managers may cause worse fund performance.

Table 2: The effect of fund network centrality on investment performance.

	Performance		
	(1-1)	(1-2)	(1-3)
Closeness	12.783**		
	(5.985)		
Betweenness		5.500	
		(5.491)	
Eigenvector			2.059*
			(1.136)
Netasset	−0.139***	−0.136***	−0.139***
	(0.029)	(0.029)	(0.030)
Familysize	−0.146**	−0.145**	−0.146**
	(0.063)	(0.063)	(0.063)
Fundage	0.353***	0.356***	0.352***
	(0.067)	(0.067)	(0.067)
Expense	1.900	1.939	1.880
	(1.223)	(1.224)	(1.222)
Turnover	−0.106***	−0.106***	−0.105***
	(0.015)	(0.015)	(0.015)
Return	−0.007	−0.006	−0.007
	(0.006)	(0.006)	(0.006)
Fundflow	0.068	0.067	0.068
	(0.042)	(0.042)	(0.042)
PhD	0.059	0.057	0.058
	(0.065)	(0.065)	(0.065)
Gender	0.003	0.003	0.003
	(0.066)	(0.066)	(0.066)
Teamsize	−0.054	−0.052	−0.053
	(0.038)	(0.038)	(0.038)
Tenure	−0.083***	−0.083***	−0.083***
	(0.027)	(0.027)	(0.027)
Time fixed effect	Included	Included	Included
Style fixed effect	Included	Included	Included
Constant	4.517***	6.026***	6.060***
	(1.512)	(1.373)	(1.366)
Cluster	Fund	Fund	Fund
Observation	37607	37607	37607
R^2 (*within*)	0.211	0.211	0.211

Notes: standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

7. Hard-to-Value Portfolio and Investment Performance

7.1. Model Specification. The fixed effect model used to test the relationships between hard-to-value portfolio and investment performance is set as follows:

$$Performance_{j,t} = \alpha + \beta \cdot HTV\_mrkt_{j,t} + \sum_i \lambda_i Control_{ij,t-1} + \varepsilon_{j,t}, \quad (11)$$

where $Performance_{j,t}$ represents the investment performance of fund j in quarter t and $HTV\_mrkt_{j,t}$ represents the hard-to-value portfolio caused by market information uncertainty of fund j's in quarter t, which can be replaced by $HTV\_fdmt_{j,t}$ or $HTV\_info_{j,t}$ when the return of hard-to-value portfolio caused by fundamental information uncertainty or limited public information is tested. The other control variables of equation (11) are those that included in equation (10). $\varepsilon_{i,t}$ denotes the error term.

7.2. Results. The regression results are shown in Table 3. Models (2-1) and (2-2) show that the parameter estimate for *HTV_mrkt* and *HTV_fdmt* are both positive ($\beta > 0$) and are significant at the 1% level and the 5% level, respectively. According to the results, mutual funds holding hard-to-value portfolios caused by market or fundamental information uncertainty will achieve better investment performance. Guo et al. [37] points out that the market or fundamental information uncertainty can lead to asset mispricing [81]. To this end, mutual funds have some capabilities of information mining and analysis to deal with that uncertain information, thus taking use of the mispricing of hard-to-value stocks in their portfolios to improve investment performance.

Table 3: The effect of hard-to-value portfolio on investment performance.

	Performance		
	(2-1)	(2-2)	(2-3)
HTV_mrkt	0.236***		
	(0.047)		
HTV_fdmt		0.046**	
		(0.021)	
HTV_info			−0.175*
			(0.110)
Netasset	−0.135***	−0.137***	−0.138***
	(0.029)	(0.029)	(0.029)
Familysize	−0.146**	−0.147**	−0.146**
	(0.063)	(0.063)	(0.063)
Fundage	0.357***	0.355***	0.353***
	(0.067)	(0.067)	(0.067)
Expense	1.967	1.932	1.897
	(1.223)	(1.223)	(1.222)
Turnover	−0.107***	−0.106***	−0.105***
	(0.015)	(0.015)	(0.015)
Return	−0.006	−0.006	−0.007
	(0.006)	(0.006)	(0.006)
Fundflow	0.069	0.068	0.067
	(0.042)	(0.042)	(0.042)
PhD	0.059	0.059	0.058
	(0.065)	(0.065)	(0.065)
Gender	0.009	0.005	<0.001
	(0.066)	(0.066)	(0.067)
Teamsize	−0.052	−0.052	−0.052
	(0.038)	(0.038)	(0.038)
Tenure	−0.085***	−0.084***	−0.081***
	(0.027)	(0.027)	(0.027)
Time fixed effect	Included	Included	Included
Style fixed effect	Included	Included	Included
Constant	5.908***	6.119***	6.230***
	(1.365)	(1.366)	(1.370)
Cluster	Fund	Fund	Fund
Observation	37607	37607	37607
R2(*within*)	0.211	0.211	0.211

Notes: standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Model (2-3) shows that effect of *HTV_info* on *Performance* is significantly negative ($\beta < 0$; $p < 0.1$), indicating that holding hard-to-value portfolio caused by limited public information will make fund performance worse. Poor disclosure of public information negatively affects the investment research of mutual funds [27]. Although information capabilities of mutual funds can help them deal with information uncertainty from market and company fundamentals, they may become helpless when the public information that can be used in investment decision is limited. That is why mutual funds holding hard-to-value portfolios shaped by poor disclosure of public information may lead to bad investment performance.

To sum up, although the empirical results do not fully support Hypothesis 2 because holding a hard-to-value portfolio caused by limited public information has negative influences on fund performance, they indicate that there are different returns between hard-to-value portfolios caused by the market or fundamental information uncertainty and hard-to-value portfolios caused by limited public information.

8. Fund Network Centrality and Return of Hard-to-Value Stock Portfolio

8.1. Model Specification. To test the impacts of fund network centrality on mutual fund's return from hard-to-value portfolio, the fixed effect model is used as follows:

$$\begin{aligned} Performance_{j,t} = {} & \alpha + \beta_1 \cdot Closeness_{j,t-1} + \beta_2 \cdot HTV\_mrkt_{j,t} \\ & + \beta_3 \cdot Closeness_{j,t-1} \cdot HTV\_mrkt_{j,t} \\ & + \sum_i \lambda_i Control_{ij,t-1} + \varepsilon_{j,t}, \end{aligned} \tag{12}$$

where $Performance_{j,t}$ represents the investment performance of fund j in quarter t, $Closeness_{j,t-1}$ represents the closeness centrality of fund j in quarter t–1, and $HTV\_mrkt_{j,t}$ represents the hard-to-value portfolio caused by market information uncertainty of fund j in quarter t. The interaction term $Closeness_{j,t-1} \times HTV\_mrkt_{j,t}$ is included to examine the impact of closeness centrality on the return of hard-to-value portfolio caused by market information uncertainty. In order to capture the effects of the three network centrality indicators on the returns of the three types of hard-to-value portfolio, $Closeness_{j,t-1}$ can be replaced by $Betweenness_{j,t-1}$ or $Eigenvector_{j,t-1}$ and $HTV\_mrkt_{j,t}$ can be replaced by $HTV\_fdmt_{j,t}$ or $HTV\_info_{j,t}$. To avoid multicollinearity, any two variables used to construct interaction terms are standardized before their multiplication. The other control variables of (12) are those that are included in equation (10). $\varepsilon_{i,t}$ denotes the error term.

8.2. Results. The regression results are shown in Table 4. Models (3-4) to (3-6) show that all interaction terms between *Betweenness* and *HTV_mrkt*, *HTV_fdmt*, or *HTV_info* are not significant at the 10% confidence level. These results further indicate that betweenness centrality cannot provide a mutual fund with information advantage that can be used in investment activities, making it unable to help the mutual fund profit from its hard-to-value portfolios.

Model (3-3) shows that the coefficient of *Closeness* × *HTV_info* is significantly positive ($\beta > 0$; $p < 0.1$), indicating that mutual funds with high closeness centrality are more likely to avoid loss from holding hard-to-value portfolios caused by limited public information. High closeness centrality can improve the breadth of the information acquired by a mutual fund and enable the fund to access to richer information sources [41]. Therefore, mutual funds with high closeness centrality can break through the limitations of poor disclosure of public information. The coefficient of *Eigenvector* × *HTV_info* is significantly positive in Model (3-9) as well ($\beta > 0$; $p < 0.01$), indicating that eigenvector centrality can also reduce the negative impacts of holding hard-to-value portfolios caused by limited public information. High eigenvector centrality can improve a mutual fund's depth of information acquisition, especially

Table 4: The effect of fund network centrality on the performance of hard-to-value portfolio.

	Performance								
	(3-1)	(3-2)	(3-3)	(3-4)	(3-5)	(3-6)	(3-7)	(3-8)	(3-9)
Closeness × HTV_mrkt	0.352 (2.304)								
Closeness × HTV_fdmt		0.462 (1.160)							
Closeness × HTV_info			9.318* (5.841)						
Betweenness × HTV_mrkt				7.049 (9.064)					
Betweenness × HTV_fdmt					0.215 (4.102)				
Betweenness × HTV_info						0.916 (23.508)			
Eigenvector × HTV_mrkt							1.931 (1.701)		
Eigenvector × HTV_fdmt								1.541** (0.709)	
Eigenvector × HTV_info									15.611*** (4.342)
Closeness	15.600** (6.536)	12.562* (6.770)	10.108 (6.212)						
Betweenness				0.824 (11.304)	6.933 (12.461)	4.319 (10.427)			
Eigenvector							0.428 (2.327)	−2.284 (2.425)	−4.203** (2.042)
HTV_mrkt	0.239*** (0.061)			0.229*** (0.050)			0.217*** (0.055)		
HTV_fdmt		0.041 (0.028)			0.049** (0.022)			0.022 (0.025)	
HTV_info			−0.292*∗ (0.145)			−0.171 (0.120)			−0.354*** (0.128)
Netasset	−0.137*** (0.029)	−0.139*** (0.029)	−0.139*** (0.029)	−0.132*** (0.029)	−0.135*** (0.029)	−0.137*** (0.030)	−0.135*** (0.029)	−0.137*** (0.029)	−0.137*** (0.029)
Familysize	−0.146** (0.063)	−0.147** (0.063)	−0.145** (0.063)	0.145** (0.063)	−0.146** (0.063)	−0.146** (0.063)	−0.146** (0.063)	0.148** (0.063)	−0.145** (0.063)
Fundage	0.354*** (0.067)	0.353*** (0.067)	0.351*** (0.067)	0.358*** (0.067)	0.357*** (0.067)	0.354*** (0.067)	0.354*** (0.067)	0.355*** (0.067)	0.353*** (0.067)
Expense	1.937 (1.223)	1.908 (1.223)	1.875 (1.223)	2.002 (1.226)	1.952 (1.224)	1.910 (1.224)	1.914 (1.222)	1.901 (1.222)	1.842 (1.223)
Turnover	−0.106*** (0.015)	−0.106*** (0.015)	−0.105*** (0.015)	−0.107*** (0.015)	−0.106*** (0.015)	−0.106*** (0.015)	−0.106*** (0.015)	−0.106*** (0.015)	−0.105*** (0.015)
Return	−0.006 (0.006)	−0.006 (0.006)	−0.007 (0.006)	−0.005 (0.006)	−0.006 (0.006)	−0.007 (0.006)	−0.006 (0.006)	−0.007 (0.006)	−0.008 (0.006)

Table 4: Continued.

	Performance								
	(3-1)	(3-2)	(3-3)	(3-4)	(3-5)	(3-6)	(3-7)	(3-8)	(3-9)
Fundflow	0.071*	0.070*	0.069*	0.070*	0.069	0.067	0.071*	0.070*	0.066
	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
PhD	0.061	0.060	0.058	0.058	0.058	0.057	0.059	0.058	0.058
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
Gender	0.009	0.005	<0.001	0.010	0.005	0.001	0.008	0.004	<−0.001
	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)
Teamsize	−0.054	−0.054	−0.053	−0.051	−0.052	−0.052	−0.053	−0.052	−0.052
	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)
Tenure	−0.085***	−0.084***	−0.082***	−0.085***	−0.084***	−0.081***	−0.085***	−0.084***	−0.081***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Time fixed effect	Included	Included	Included	Included	Included	Included	Included	Included	Included
Style fixed effect	Included	Included	Included	Included	Included	Included	Included	Included	Included
Constant	3.899**	4.479***	4.739***	5.786***	5.972***	6.134***	5.893***	6.248***	6.400***
	(1.516)	(1.533)	(1.518)	(1.371)	(1.375)	(1.375)	(1.361)	(1.364)	(1.361)
Cluster	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund	Fund
Observation	37607	37607	37607	37607	37607	37607	37607	37607	37607
R2(*within*)	0.212	0.211	0.211	0.211	0.211	0.211	0.212	0.211	0.211

Notes: standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

helping the mutual fund obtain potential private information from other entities having important information positions. More private information can supplement the lack of public information when the mutual fund analyzes hard-to-value assets caused by limited public information.

In Model (3-8), the coefficient of *Eigenvector* × *HTV_fdmt* is significantly positive ($\beta > 0$; $p < 0.05$), while the coefficient of *Closeness* × *HTV_fdmt* is not significant at the 10% confidence level. The results indicate that the advantages of fundamental information rely more on the information depth than the information breadth. More specifically, high-quality information from entities having important information positions (i.e., high eigenvector centrality) can help mutual funds gain returns from hard-to-value portfolios caused by fundamental information uncertainty, while obtaining public and private information through high-efficiency information transmission path (i.e., high closeness centrality) contributes little to the performance of this type of hard-to-value portfolio.

The coefficients of *Closeness* × *HTV_mrkt* and *Eigenvector* × *HTV_mrkt* both are not significant at the 10% confidence level. The results indicate that high closeness centrality or high eigenvector centrality cannot help mutual funds achieve better performance from holding hard-to-value portfolios caused by market information uncertainty. In other words, it is quite difficult for a mutual fund to use its information breadth and depth brought by the fund network to deal with the information uncertainty at the market level. The possible reason may be that information about stock price spread more easily than information about company fundamental, so the information exchanges in the fund network are more likely to make mutual funds have similar market information rather than fundamental information. Consequently, homogeneous market information cannot guarantee a mutual fund to obtain enough market information advantages relative to others by using its network position, thus preventing it to profit from its hard-to-value portfolios caused by market information uncertainty.

In summary, the results support Hypothesis 3 to some extent, indicating that the information breadth brought by closeness centrality have positive influences on the returns of hard-to-value portfolios caused by limited public information, and the information depth brought by eigenvector centrality have positive influences both on the returns of hard-to-value portfolios caused by limited public information or fundamental information uncertainty.

9. Conclusions

Using the quarterly data of China's mutual funds from the fourth quarter of 2004 to the fourth quarter of 2019, this paper constructs a series of complex bipartite networks based on the overlapped portfolios of mutual funds and then explores the mechanism by how the position of a mutual fund in the fund network influences its performance. The results show that the information breadth brought by closeness centrality and the information depth brought by eigenvector centrality help a mutual fund achieve better investment performance, but the information intermediation brought by betweenness centrality has no significant influence on investment performance. We also find that holding stock portfolios with high valuation difficulties caused by the market or fundamental information uncertainty will lead to better fund performance, but while holding hard-to-value portfolios caused by limited public information has negative influences on fund performance. Furthermore, closeness centrality and eigenvector centrality can reduce the likelihood of a mutual fund holding hard-to-value portfolios caused by limited public information to achieve worse performance. Eigenvector centrality positively moderates the relationship between hard-to-value portfolios caused by fundamental information uncertainty and investment performance.

The main contributions of this paper are as follows. First, most of the previous studies use the degree centrality of a one-mode fund network to explore the influences of the fund network position [12, 24, 59]. Based on graph theory, we construct bipartite networks based on the overlapped portfolios of mutual funds and calculate the closeness centrality, betweenness centrality, and eigenvector centrality of mutual funds in the networks, and then reveal the influences of the three centrality indicators on mutual fund's investment behavior and performance. Compared with a one-mode fund network, a bipartite network can comprehensively describe the information channels of mutual funds to access other funds and listed companies. Besides, closeness centrality, betweenness centrality, and eigenvector centrality reflect the overall picture of all direct and indirect relations and capture richer structural features of the fund network compared to degree centrality. At the same time, we further indicate that the three centrality indicators bring different information advantages, so they have different influences on mutual funds. Second, this paper constructs indicators to measure three types of hard-to-value portfolios according to the three causes (i.e., market information uncertainty, fundamental information uncertainty, and limited public information). The three types of hard-to-value portfolios have different information natures, thus bringing different influences on fund performance. Third, this paper examines the effects of closeness centrality, betweenness centrality, and eigenvector centrality on the returns of hard-to-value portfolios, respectively. We emphasize that a mutual fund should make appropriate use of the information advantages brought by the fund networks to improve its capabilities of information mining and analysis required for investing in hard-to-value stocks.

There are two main practical implications. First, for the investors of mutual funds, the methods in the paper about constructing the bipartite network based on the overlapped portfolios of mutual funds and calculating the fund network centrality indicators can help them judge the information advantages of each mutual fund. In fact, closeness centrality and eigenvector centrality are two important drives for fund performance. Second, for fund managers, we point out that they need to expand their information breadth by establishing relations with more listed companies and other mutual funds and increase their information depth by connecting with those institutions having important

information positions. Additionally, fund managers should actively use their information advantages to invest in stocks with high levels of market or fundamental information uncertainty and avoid holding stocks having limited public information.

This paper has some limitations that can also provide promising directions for future research. First, this paper discusses the impacts of fund networks on mutual fund's investment behavior and performance, but only focuses on the analysis of network centrality. However, there are many other important network characteristics, such as network density, tie strength, and network constrain. It would be valuable to conduct more research on how other characteristics of the fund network perform. Second, we explore the influences of the information advantages brought by fund networks on mutual fund behavior of holding hard-to-value portfolios. Other investment behaviors such as diversification and timing choice require more explorations in further studies.

10. Endnotes

According to Statman [82], diversified investment in more than 20 stocks can basically eliminate most of the nonsystemic risks. When a mutual fund holds more than 5% of its all position in a stock, it means that the fund manager is relatively confident in that stock and may own the private information of that stock. Therefore, we set the threshold of a mutual fund's large position as 5%.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (nos. 72002175 and 72002033), Fundamental Research Funds for the Central Universities in UIBE (nos. CXTD10-05, CXTD 11-04, and 19YQ08), Foundation for Disciplinary Development of SITM in UIBE, and Fundamental Research Funds for the Central Universities in BFSU (no. 2020JJ006)

References

[1] R. Wermers, "Mutual fund performance: an empirical decomposition into stock-picking talent, style, transactions costs, and expenses," *The Journal of Finance*, vol. 55, no. 4, pp. 1655–1695, 2000.

[2] S. Gibson, A. Safieddine, and R. Sonti, "Smart investments by smart money: evidence from seasoned equity offerings," *Journal of Financial Economics*, vol. 72, no. 3, pp. 581–604, 2004.

[3] M. Baker, L. Litov, J. A. Wachter, and J. Wurgler, "Can mutual fund managers pick stocks? Evidence from their trades prior to earnings announcements," *Journal of Financial and Quantitative Analysis*, vol. 45, no. 5, pp. 1111–1131, 2010.

[4] M. Kacperczyk, S. V. Nieuwerburgh, and L. Veldkamp, "Time-varying fund manager skill," *The Journal of Finance*, vol. 69, no. 4, pp. 1455–1484, 2014.

[5] M. Granovetter, "Economic action and social structure: the problem of embeddedness," *American Journal of Sociology*, vol. 91, no. 3, pp. 481–510, 1985.

[6] J. Y. Zhao, Y. J. Wang, X. Xi, and G. D. Wu, "Simulation of steel production logistics system based on multi-agents," *International Journal of Simulation Modelling*, vol. 16, no. 1, pp. 167–175, 2017.

[7] B. Andres, R. Poler, L. M. Camarinha-Matos, and H. Afsarmanesh, "A simulation approach to assess partners selected for a collaborative network," *International Journal of Simulation Modelling*, vol. 16, no. 3, pp. 399–411, 2017.

[8] H. S. Lee and W. S. Lee, "International linkage among mena financial markets," *Economic Computation & Economic Cybernetics Studies & Research*, vol. 53, no. 3, pp. 277–296, 2019.

[9] R. S. Burt, *Structural Holes: The Social Structure of Competition*, Harvard University Press, Cambridge, MA, USA, 1992.

[10] J. M. Podolny, "Networks as the pipes and prisms of the market," *American Journal of Sociology*, vol. 107, no. 1, pp. 33–60, 2001.

[11] H. Hong, J. D. Kubik, and J. C. Stein, "Thy neighbor's portfolio: word-of-mouth effects in the holdings and trades of money managers," *The Journal of Finance*, vol. 60, no. 6, pp. 2801–2824, 2005.

[12] A. Pareek, "Information networks: implications for mutual fund trading behavior and stock returns," in *in Proceedings of the AFA 2010 Atlanta Meetings Paper*, Atlanta, GA, USA, January 2012.

[13] V. K. Pool, N. Stoffman, and S. E. Yonker, "The people in your neighborhood: social interactions and mutual fund portfolios," *The Journal of Finance*, vol. 70, no. 6, pp. 2679–2732, 2015.

[14] B. J. Bushee and T. H. Goodman, "Which institutional investors trade based on private information about earnings and returns?" *Journal of Accounting Research*, vol. 45, no. 2, pp. 289–321, 2007.

[15] H. Jiang, "Institutional investors, intangible information, and the book-to-market effect," *Journal of Financial Economics*, vol. 96, no. 1, pp. 98–126, 2010.

[16] S. S. Crawford, W. R. Gray, and A. E. Kern, "Why do fund managers identify and share profitable ideas?" *Journal of Financial and Quantitative Analysis*, vol. 52, no. 5, pp. 1903–1926, 2017.

[17] J. L. Iribarren and E. Moro, "Affinity paths and information diffusion in social networks," *Social Networks*, vol. 33, no. 2, pp. 134–142, 2011.

[18] P.-H. Soh and E. B. Roberts, "Technology alliances and networks: an external link to research capability," *IEEE Transactions on Engineering Management*, vol. 52, no. 4, pp. 419–428, 2005.

[19] M. A. Carpenter, M. Li, and H. Jiang, "Social network research in organizational contexts," *Journal of Management*, vol. 38, no. 4, pp. 1328–1361, 2012.

[20] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[21] L. C. Freeman, D. Roeder, and R. R. Mulholland, "Centrality in social networks: ii. experimental results," *Social Networks*, vol. 2, no. 2, pp. 119–141, 1979.

[22] N. E. Friedkin, "Theoretical foundations for centrality measures," *American Journal of Sociology*, vol. 96, no. 6, pp. 1478–1504, 1991.

[23] Z. L. Tian, R. H. Liu, and Y. Liu, "Information transmission, collective stampede and systematic tail risk," *China Economic Quarterly*, vol. 18, no. 3, pp. 897–918, 2019, in Chinese.

[24] R. H. Luo and Z. L. Tian, "Mutual fund network, competition barrier and stock information environment," *China Industrial Economics*, vol. 3, pp. 137–154, 2020, in Chinese.

[25] Y. Shen, J. M. Zhao, and X. He, "Alumni network, funds' performance and 'Small world' effect," *China Economic Quarterly*, vol. 15, no. 1, pp. 403–428, 2016, in Chinese.

[26] J. Wurgler and E. Zhuravskaya, "Does arbitrage flatten demand curves for stocks?" *The Journal of Business*, vol. 75, no. 4, pp. 583–608, 2002.

[27] B. Cornell, W. R. Landsman, and S. R. Stubben, "Accounting information, investor sentiment, and market pricing," *Journal of Law, Finance, and Accounting*, vol. 2, no. 2, pp. 325–345, 2017.

[28] K. Chan and A. Hameed, "Stock price synchronicity and analyst coverage in emerging markets," *Journal of Financial Economics*, vol. 80, no. 1, pp. 115–147, 2006.

[29] K. Daniel, D. Hirshleifer, and A. Subrahmanyam, "Investor psychology and security market under- and overreactions," *The Journal of Finance*, vol. 53, no. 6, pp. 1839–1885, 1998.

[30] K. D. Daniel, D. Hirshleifer, and A. Subrahmanyam, "Overconfidence, arbitrage, and equilibrium asset pricing," *The Journal of Finance*, vol. 56, no. 3, pp. 921–965, 2001.

[31] D. Hirshleifer, "Investor psychology and asset pricing," *The Journal of Finance*, vol. 56, no. 4, pp. 1533–1597, 2001.

[32] B. M. Barber, T. Odean, and N. Zhu, "Do retail trades move markets?" *Review of Financial Studies*, vol. 22, no. 1, pp. 151–186, 2009.

[33] A. Kumar, "Hard-to-value stocks, behavioral biases, and informed trading," *Journal of Financial and Quantitative Analysis*, vol. 44, no. 6, pp. 1375–1401, 2009.

[34] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, vol. 61, no. 4, pp. 1645–1680, 2006.

[35] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *Journal of Economic Perspectives*, vol. 21, no. 2, pp. 129–151, 2007.

[36] S. Lai, L. Ng, and B. Zhang, "Does PIN affect equity prices around the world?" *Journal of Financial Economics*, vol. 114, no. 1, pp. 178–195, 2014.

[37] Q. Y. Guo, R. H. Luo, and Y. Liu, "Difficulty in stock analysis and mutual fund's information advantage," *Chinese Review of Financial Studies*, vol. 9, no. 6, pp. 16–32, 2017, in Chinese.

[38] W. W. Powell, K. W. Koput, and L. Smith-Doerr, "Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology," *Administrative Science Quarterly*, vol. 41, no. 1, pp. 116–145, 1996.

[39] D. J. Brass and M. E. Burkhardt, "Potential power and power use: an investigation of structure and behavior," *Academy of Management Journal*, vol. 36, no. 3, pp. 441–470, 1993.

[40] T. J. Rowley, "Moving beyond dyadic ties: a network theory of stakeholder influences," *Academy of Management Review*, vol. 22, no. 4, pp. 887–910, 1997.

[41] R. Gulati, "Network location and learning: the influence of network resources and firm capabilities on alliance formation," *Strategic Management Journal*, vol. 20, no. 5, pp. 397–420, 1999.

[42] O. Sorenson and T. E. Stuart, "Syndication networks and the spatial distribution of venture capital investments," *American Journal of Sociology*, vol. 106, no. 6, pp. 1546–1588, 2001.

[43] P. Hage and F. Harary, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, no. 1, pp. 57–63, 1995.

[44] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[45] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *The Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[46] H.-L. Chen, N. Jegadeesh, and R. Wermers, "The value of active mutual fund management: an examination of the stockholdings and trades of fund managers," *The Journal of Financial and Quantitative Analysis*, vol. 35, no. 3, pp. 343–368, 2000.

[47] J. P. H. Fan and T. J. Wong, "Corporate ownership structure and the informativeness of accounting earnings in East Asia," *Journal of Accounting and Economics*, vol. 33, no. 3, pp. 401–425, 2002.

[48] R. Ding, W. Hou, J.-M. Kuo, and E. Lee, "Fund ownership and stock price informativeness of Chinese listed firms," *Journal of Multinational Financial Management*, vol. 23, no. 3, pp. 166–185, 2013.

[49] D. Liu, H. Gu, and P. Lung, "The equity mispricing: evidence from China's stock market," *Pacific-Basin Finance Journal*, vol. 39, no. 39, pp. 211–223, 2016.

[50] W. M. Fong, "Speculative trading and stock returns: a stochastic dominance analysis of the Chinese A-share market," *Journal of International Financial Markets, Institutions and Money*, vol. 19, no. 4, pp. 712–727, 2009.

[51] E. Xiang, G. Y. Tian, F. Yang, and Z. Liu, "Do mutual funds have information advantage? Evidence from seasoned equity offerings in China," *International Review of Financial Analysis*, vol. 31, pp. 70–79, 2014.

[52] R. B. Cohen, C. Polk, and B. Silli, "Best ideas," London School of Economics, London, UK, 2010.

[53] R. Shiller and J. Pound, "Survey evidence of diffusion of interest among institutional investors," *NBER Working Paper 1851*, 1986.

[54] B. J. Bushee, M. J. Jung, and G. S. Miller, "Conference presentations and the disclosure milieu," *Journal of Accounting Research*, vol. 49, no. 5, pp. 1163–1192, 2011.

[55] B. J. Bushee, J. Gerakos, and L. F. Lee, "Corporate jets and private meetings with investors," *Journal of Accounting and Economics*, vol. 65, no. 2-3, pp. 358–379, 2018.

[56] L. Cohen, A. Frazzini, and C. Malloy, "The small world of investing: board connections and mutual fund returns," *Journal of Political Economy*, vol. 116, no. 5, pp. 951–979, 2008.

[57] R. Frankel, M. Johnson, and D. J. Skinner, "An empirical examination of conference calls as a voluntary disclosure medium," *Journal of Accounting Research*, vol. 37, no. 1, pp. 133–150, 1999.

[58] D. Delpini, S. Battiston, G. Caldarelli, and M. Riccaboni, "The network of U.S. mutual fund investments: diversification, similarity and fragility throughout the global financial crisis," 2018, https://arxiv.org/abs/1801.02205.

[59] J. F. Lavin, M. A. Valle, and N. S. Magner, "Modeling overlapped mutual funds' portfolios: a bipartite network spproach," *Complexity*, vol. 2019, Article ID 1565698, 20 pages, 2019.

[60] V. Agarwal, K. A. Mullally, Y. Tang, and B. Yang, "Mandatory portfolio disclosure, stock liquidity, and mutual fund performance," *The Journal of Finance*, vol. 70, no. 6, pp. 2733–2776, 2015.

[61] M. M. Carhart, "On persistence in mutual fund performance," *The Journal of Finance*, vol. 52, no. 1, pp. 57–82, 1997.
[62] K. Faust, "Centrality in affiliation networks," *Social Networks*, vol. 19, no. 2, pp. 157–191, 1997.
[63] P. Bonacich, "Communication networks and collective action," *Social Networks*, vol. 9, no. 4, pp. 389–396, 1987.
[64] Y. Amihud, "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, vol. 5, no. 1, pp. 31–56, 2002.
[65] W. Kang, N. Li, and H. Zhang, "Information uncertainty and the pricing of liquidity," *Journal of Empirical Finance*, vol. 54, pp. 77–96, 2019.
[66] O. A. Lamont and R. H. Thaler, "Can the market add and subtract? Mispricing in tech stock carve-outs," *Journal of Political Economy*, vol. 111, no. 2, pp. 227–268, 2003.
[67] R. La Porta, F. Lopez-de-Silanes, A. Shleifer, and R. Vishny, "Law and finance," *Journal of Political Economy*, vol. 106, no. 6, pp. 1113–1155, 1998.
[68] M. H. Lang and R. J. Lundholm, "Corporate disclosure policy and analyst behavior," *Accounting Review*, vol. 71, no. 4, pp. 467–492, 1996.
[69] H. Hong, T. Lim, and J. C. Stein, "Bad news travels slowly: size, analyst coverage, and the profitability of momentum strategies," *The Journal of Finance*, vol. 55, no. 1, pp. 265–295, 2000.
[70] C. A. Gleason and C. M. C. Lee, "Analyst forecast revisions and market price discovery," *The Accounting Review*, vol. 78, no. 1, pp. 193–225, 2003.
[71] M. Simutin, "Standing out in the fund family: deviation from a family portfolio predicts mutual fund performance," University of Toronto, Toronto, Canada, 2013.
[72] A. Niessen-Ruenzi and S. Ruenzi, "Sex matters: gender bias in the mutual fund industry," *Management Science*, vol. 65, no. 7, pp. 2947–3448, 2018.
[73] M. Kacperczyk and A. Seru, "Fund manager use of public information: new evidence on managerial skills," *The Journal of Finance*, vol. 62, no. 2, pp. 485–528, 2007.
[74] K. J. M. Cremers and A. Petajisto, "How active is your fund manager? A new measure that predicts performance," *Review of Financial Studies*, vol. 22, no. 9, pp. 3329–3365, 2009.
[75] Y. Amihud and R. Goyenko, "Mutual Fund'sR2as predictor of performance," *Review of Financial Studies*, vol. 26, no. 3, pp. 667–694, 2013.
[76] K. Daniel, M. Grinblatt, S. Titman, and R. Wermers, "Measuring mutual fund performance with characteristic-based benchmarks," *The Journal of Finance*, vol. 52, no. 3, pp. 1035–1058, 1997.
[77] G. Hoberg, N. Kumar, and N. Prabhala, "Mutual fund competition, managerial skill, and alpha persistence," *The Review of Financial Studies*, vol. 31, no. 5, pp. 1896–1929, 2018.
[78] B. Mullen, C. Johnson, and E. Salas, "Effects of communication network structure: components of positional centrality," *Social Networks*, vol. 13, no. 2, pp. 169–185, 1991.
[79] J. M. Podolny, "A status-based model of market competition," *American Journal of Sociology*, vol. 98, no. 4, pp. 829–872, 1993.
[80] J. M. Podolny and D. J. Phillips, "The dynamics of organizational status," *Industrial and Corporate Change*, vol. 5, no. 2, pp. 453–471, 1996.
[81] X. F. Zhang, "Information uncertainty and stock returns," *The Journal of Finance*, vol. 61, no. 1, pp. 105–137, 2006.
[82] M. Statman, "How many stocks make a diversified portfolio?" *The Journal of Financial and Quantitative Analysis*, vol. 22, no. 3, pp. 353–363, 1987.

Hindawi
Complexity
Volume 2020, Article ID 6683834, 10 pages
https://doi.org/10.1155/2020/6683834

Research Article

A Smart Privacy-Preserving Learning Method by Fake Gradients to Protect Users Items in Recommender Systems

Guixun Luo,[1] **Zhiyuan Zhang**,[2] **Zhenjiang Zhang**,[3] **Yun Liu**,[2] **and Lifu Wang**[2]

[1]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China*
[3]*School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Zhiyuan Zhang; zhangzhiyuan@bjtu.edu.cn

Received 22 October 2020; Revised 2 November 2020; Accepted 27 November 2020; Published 17 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

In this paper, we study the problem of protecting privacy in recommender systems. We focus on protecting the items rated by users and propose a novel privacy-preserving matrix factorization algorithm. In our algorithm, the user will submit a fake gradient to make the central server not able to distinguish which items are selected by the user. We make the Kullback–Leibler distance between the real and fake gradient distributions to be small thus hard to be distinguished. Using theories and experiments, we show that our algorithm can be reduced to a time-delay SGD, which can be proved to have a good convergence so that the accuracy will not decline. Our algorithm achieves a good tradeoff between the privacy and accuracy.

1. Introduction

Recommender systems, which help the electronic commerce websites to give more useful suggestions, are becoming more and more important. However, to provide users with appropriate options, the server will collect users' data, which includes lots of sensitive information.

Data in electronic commerce, economics, supply chain, financial system [1–10], etc., are generally very sensitive. In the electronic commerce case, it is shown in many studies, such as [11, 12] that user data in recommender systems, shopping records, movies a user has watched, and ratings for the restaurants contain lots of very private information such as political attitudes, sexual orientation, etc. In this paper, we study the privacy-protecting problem in electronic commerce data. Privacy has been an important issue for a long time, not only in the recommender system but also in almost all algorithms in data mining and machine learning.

Differential privacy [13] is a popular method to protect privacy in machine learning algorithms. For recommender systems, there are many works applying differential privacy, such as [14–16]. Differential privacy matrix factorization algorithms are introduced in [17, 18], etc. Traditional differential privacy method is centralized, in other words, relying on a trustworthy data collector. When we want the central server not to be able to get privacy information, local differential privacy (LDP) should be used. Every user will add noise to their private data in their own device before being submitted to the central server. Recommender systems with LDP are studied in [19–21]. LDP has been used in Google's Chrome browser [22] and Apple IOS 10 [23] to collect user data.

In local differential privacy, there are two important things to be protected. The first one is which items this user has rated and the second one is the ratings of the user. In some situations, which items have been rated is much more sensitive than the rating itself. For example, shopping record contains a lot of private information, but the ratings can only represent the quality of goods. The work in [19] can only protect the ratings but not both. Shin et al. [17] proposed a novel LDP matrix factorization algorithm to protect both kinds of privacy information based on the work in [24]. Their method is to let the user submit a noisy gradient, whose value is either B or $-B$. The

algorithm is ε-LDP, and in each round of the training process, and since the output is binary, the adversary can not learn about which items are rated in a single iteration process.

However, if the adversary can get noisy gradients in multiple iterations since the noisy gradients obey the Bernoulli distribution with a mean 0, the items which have not been rated can be identified by a statistical test. The intensity of the privacy protection for the ratings and items after multiple iterations can be guaranteed by composition theorems for LDP [25, 26]. If every iteration is ε-LDP, after k iterations, the final algorithm is at most $k\varepsilon$-LDP. But these analyses are not a direct guarantee to protect the items rated by the users. We can turn to a new perspective on this question. After performing k iterations, given a sequence with length k denoted by y_i, where y_j^i is the gradient submitted in iteration i, let $P_{\text{real}}(y_j)$ be the probability that y_j is a real gradient sequence and let $P_{\text{fake}}(y_j)$ be the probability that y_j is fake. Using these two probabilities, we can consider testing two hypotheses, the sequence is real and the sequence is fake. So now comes the question, how can we make it difficult to distinguish the two situations?

In order to improve the ability of protecting privacy, we want the probability error to be large. Note that the average negative log probability of error is well-known deduced from the Chernoff–Stein lemma.

Theorem 1 (Theorem 11.8.3 in [27]). *$X \sim Q$ is a random variable; consider the hypothesis test between two alternatives, $Q = P_1$ and $Q = P_2$, where $D(P_1, P_2)$, the K-L distance, is finite. Then the average negative log probability of error of this hypothesis testing is $D(P_1, P_2)$.*

Using this result, although we can not obtain the distribution of the real sequence, in Section 4, we will show that for the Gaussian noise based differential privacy algorithm, we can estimate the mean value of K-L distance and optimize the value of fake gradient to make the two distributions to be difficult to distinguish.

In this paper, we propose a novel algorithm such that if the item has not been rated by the user, the user will submit a fake gradient. Else, the user can submit the real one, but all the submitted data will eventually be noise added. The paper is organized as follows. In Section 2, we introduce differential privacy briefly as preliminaries. In Section 3, we introduce the framework of the general differential privacy matrix factorization algorithm. And in Section 4, we will show that our algorithm can reduce the average K-L distance between the fake and real gradient distributions, such that it can improve the intensity to protect the privacy items. Meanwhile, we can prove that our algorithm has the form of SGD with time delay, which can be proved that the accuracy of the model will not be reduced by our updating rules so that our algorithm achieves a tradeoff between accuracy and privacy. In Section 5, we use experiments to show the effectiveness of our algorithm. The related work is reviewed in Section 6. In the final section, we conclude.

Table 1: Notations.

Notation	Meaning
m	Number of users
n	Number of items
u	The user profile vector
v	The item profile vector
$L(u, v)$	The loss function
M	The set of the ratings
C	Bound of the norm of the gradient in privacy gradient descent
η	The learning rate
λ	The regularization parameters
β	β-Smooth parpameters for the loss function

2. Preliminaries

In this paper, the notations we used are listed in Table 1.

2.1. Differential Privacy. Differential Privacy is first introduced by Dwork et al. [13], the aim of which is to make it difficult for an attacker to obtain privacy from the output data by adding noise.

Definition 1. A randomized algorithm $M: D \longrightarrow R$ with domain D and range R is (ε, δ)-Differential Privacy, if for two adjacent data $d, d' \in D$ and for a subset S of range R, it holds that

$$P(M(d) \in S) \leq e^{\varepsilon} P(M(d') \in S) + \delta. \tag{1}$$

Note that this definition is to compare the two probability. If $\delta = 0$, it can be expressed as

$$\ln \frac{P(M(d) \in S)}{P(M(d') \in S)} \leq \varepsilon. \tag{2}$$

If ε is small, such that it is hard to distinguish whether the output data is come from d or d'. As in [28], one can link differential privacy with mutual-information.

Another way to describe Differential Privacy is to use the distance between distributions. We say a randomized algorithm $M: D \longrightarrow R$ is (α, ε) Renyi Differential Privacy if for all neighboring d' and $'d'$ we have

$$D_{\alpha}(M(d'), M(d)) = \frac{1}{\alpha - 1} \ln \int \left(\frac{M(d')(z)}{M(d)(z)} \right)^{\alpha} M(d)(z) \mathrm{d}z \leq \varepsilon. \tag{3}$$

When $\alpha \longrightarrow 1$, D_1 is the Kullback-Leibler distance, and when $\alpha = \infty$, Renyi Differential Privacy is equal to $(\varepsilon, 0)$ Differential Privacy. So we can see Differential Privacy is to make the output distributions with different inputs to be indistinguishable(the distributions have small distances).

One may ask how to achieve (ε, δ)-Differential Privacy in machine learning process. A basic paradigm to achieve "-- differential privacy is to examine a query L_2-sensitivity in [29].

Definition 2. f is a map from the data in the dataset D to a vector. The L_2-sensitivity of f is $\Delta_2(f) = \max_{d', d} \| f(d) - f(d) \|$.

Using this definition, we have the following theorem in [29].

Theorem 2. *If f is a map from D to $\mathbb{R}^d$. Then the randomized algorithm $M(D)$: $f(D) + n$ where*

$$P(n) \propto \exp\left(\frac{\varepsilon \|n\|}{\Delta_2(f)}\right), \tag{4}$$

achieves ε-Differential Privacy.

This theorem provides a basic method to achieve Differential-Privacy-Machine-Learning.

3. The Framwork of Perturbed Matrix Factorization Algorithm

The program of Matrix Factorization algorithm with privacy protection has been studied by many authors, such as [17, 19].

When minimizing the cost function

$$\begin{aligned} &L(u, v) + \lambda\left(\sum_i \|u_i\|^2 + \sum_j \|v_j\|^2\right) \\ &= \frac{1}{n}\sum_{i,j\in M}\left(r_{ij} - u_i^T v_j\right)^2 + \lambda\left(\sum_i \|u_i\|^2 + \sum_j \|v_j\|^2\right). \end{aligned} \tag{5}$$

We can use gradient descent

$$\begin{aligned} u_i(t+1) &= u_i(t) - \eta\nabla_{u_i} L(u(t), v(t)) + 2\lambda u_i(t), \\ v_j(t+1) &= v_j(t) - \eta\nabla_{v_j} L(u(t), v(t)) + 2\lambda v_j(t). \end{aligned} \tag{6}$$

The vector u_i is the user profile vector for user i, and v_j is the item profile vector for item j.

Note that we have

$$\sum_{i,j\in M}\left(r_{ij} - u_i^T v_j\right)^2 = \sum_{i,j} y_{ij}\left(r_{ij} - u_i^T v_j\right)^2, \tag{7}$$

$$\nabla_{v_j} L(u, v) = -\frac{2}{n}\sum_i g_{ij} = -\frac{2}{n}\sum_i y_{ij} u_i\left(r_{ij} - u_i^T v_j\right), \tag{8}$$

where $y_{ij} = 1$ if $i, j \in M$ else. $y_{ij} = 0$

In this type of program, the user profile vectors u_i are saved and updated on the users' own devices. As for the item profile vectors, all the users will send the gradient to the central server, and individual users should perturb their gradient g_{ij} using a random mechanism $\mathbb{M}$. Then the central server sums all these gradients to update the item profile vectors v_j. Using this random perturbation, ε-differential privacy can be achieved by adjusting the distribution of noise.

The whole process is shown in Algorithm 1.

Note that there are two types of private information. One is the ratings of the users and the other one is the items have been rated by the users.

In order to protect the items, one way is to use the random response mechanism introduced in Section 4.1 of [17]. In this method, we generate a y'_{ij} such that $y'_{ij} = 1$ with probability p, and if the original $y_{ij} = 0$, we set a fake rating $r_{ij} = 0$ so the fake gradient is $u_i(0 - u_i^T v_j)$ by (8), and Gaussian noise is added to the final gradient sent to the central server to protect the ratings of users.

However, it is shown in the discussion of Section 4.1 of [17] that the error caused by these fake ratings is not small, which will influence the final model accuracy. The main reason is that there are many fake gradients, which lead to a great error in the expectation of the sum of gradients.

One way is to solve this problem is to set the fake gradient F_{ij} to be zero. If $y_{ij} = 0$, the user sends a random variable $\mathbb{M}(0)$ to the central server. This method is used in [17], where $\mathbb{M}(x)$ is a Bernoulli random variable with mean value x. However, the disadvantage of this method is that the distribution of gradients in the $y_{ij} = 0$ case is very different from the distribution of the real gradient. For example, we can collect some data of g_{ij} sent by the user i, and use a statistical test to test if this data obeys the certain distribution of mean 0, then we can know whether $y_{ij} = 0$.

All in all, we need to strike a balance between privacy and accuracy. We need to provide a fake gradient to make sure the accuracy will not be greatly affected and let these two distributions, the fake one and the real one, to be statistically indistinguishable as far as possible.

4. The Main Results

In this paper, since we are concerned about the items of users, we will focus on considering the statistical distance of $y_{ij} = 0$ and $y_{ij} = 1$ distributions. We propose a novel algorithm to protect items of the users. In our algorithm, the user will submit a noise-added fake gradient in the $y_{ij} = 0$ case. The K-L distance between the real and fake distributions will be small so that they are hard to be distinguished. On the other hand, we will study how will the fake gradients influence the model accuracy. We will show that in our algorithm, the updating rules can be reduced to a time-delay SGD, which will not influence the accuracy.

In our algorithm, the random mechanism $\mathbb{M}$ we choose is the Gaussian random mechanism, $\mathbb{M}(d) = N(d, \sigma^2)$. One of the advantages is that there is a very good composition theorem [26] which gives a much tighter estimate on the multi-iteration privacy loss for Gaussian mechanism-based differential privacy gradient descent algorithm.

Theorem 3 (Theorem 1 in [26]). *Let C be the gradient bound in privacy gradient descent, there exist two constants c_1 and c_2 such that the after k iterations, the Gaussian noisy privacy gradient descent algorithm is (ε, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma \geq c_2 C \frac{\sqrt{T\ln(1/\delta)}}{\varepsilon}. \tag{9}$$

Generally, C is chosen to be a prior bound of the gradient norm, so we do not write it in the algorithm description explicitly.

Input: Random mechanism $\mathbb{M}$, learning rate η, and redefined iteration number k
Output: Item profile matrix V
Randomly initialize $u_i(0), v_j(0)$ for all i and j.
for $t = 1, 2, 3 \ldots$ **do**
  Initialize $G_j = 0$ for all j in central server.
  for $i = 1, 2, 3, \ldots, m$ **do**
    On user i: sample j uniformly
    from$\{1, 2, 3, \ldots, n\}$.
    if $y_{ij} = 1$ **then**
      $g_{ij} = u_i(r_{ij} - u_i^T v_j)$
      $g_{ij} = g_{ij}/\max(1, (\|g_{ij}\|^2/C))$
      $g'_{ij} = \mathbb{M}(x_{ij})$
    end
    else
      Generate a fake gradient of F_{ij}.
      set $g_{ij} = F_{ij}$
      $g'_{ij} = \mathbb{M}(x_{ij})$
    end
    $G_j = G_j + x'_{ij}$ for all j.
  end
  For all j:
  $G_j = (G_j/m)$
  $v_j = v_j + \eta G_j$
 for $i = 1, 2, 3 \ldots m$ **do**
    Update u_i on a local device by gradient descent.
  end
end

Algorithm 1: Perturbed Matrix Factorization algorithm.

In the case of the Gaussian random mechanism, it is easy to calculate the K-L distance between distributions. In the following section, we will show that we can find a good choice of the fake gradient.

4.1. Estimating the K-L Distance between Two Distributions. Given a gradient sequence y_j with length k, a probability of y_j can be represented in the following form.

$$P\left(y_j = o_{1:k}\right) = \prod_{i=1}^{k} P\left(y_j^i = o_i | y_j^{i-1} = o_{i-1}\right). \tag{10}$$

Using this form we can calculate K-L distance.

Given two probability measures P_1 and P_2 in length k sequence space, we have

$$\begin{aligned} D(P_1, P_2) &= \int P_1(o_{1:k}) \sum_{i=1}^{k} \log \frac{P_1(o_i|o_{i-1})}{P_2(o_i|o_{i-1})} \mathrm{d}o \\ &= \sum_{i=1}^{k} \int P_1(o_i|o_{i-1}) P_1(o_{i-1}) \log \frac{P_1(o_i|o_{i-1})}{P_2(o_i|o_{i-1})} \mathrm{d}o_i \mathrm{d}o_{i-1} \\ &= \sum_{i=1}^{k} \int P_1(o_{i-1}) D(P_1(o_i|o_{i-1}), P_2(o_i|o_{i-1})) \mathrm{d}o_{i-1}. \end{aligned} \tag{11}$$

In each iteration, the user will sent a perturbed gradient g'_{ij} to the central server, which has the following forms:

$$g'_{ij} = \begin{cases} u_i\left(r_{ij} - u_i^T v_j\right) + N\left(0, \sigma^2\right), & \text{if } y_{ij} = 1, \\ F_{ij} + N\left(0, \sigma^2\right), & \text{if } y_{ij} = 0. \end{cases} \tag{12}$$

The we have $P_{\text{real}}(o_t|o_{t-1}) = P(x_t + \eta_{ij} = o_t)$, where x_t is the gradient calculated from o_{t-1}, so $D(P_{\text{real}}(o_t|o_{i-1}), P_{\text{fake}}(o_t|o_{i-1})) = D(N(x_t^{ij}, \sigma^2), N(G_{ij}, \sigma^2)$.

This is the K-L distance between two Gaussian distributions with the same σ. We can show that $D(N(x_t^{ij}, \sigma^2), N(G_{ij}, \sigma^2)) = \text{const} + (1/2\sigma^2)\|x_t^{ij} - G_{ij}\|^2$.

From equation (11), if we want to optimize the K-L distance, we need to consider

$$\sum_t \int P_{\text{real}}\left(x_t^{ij}\right) \frac{1}{2\sigma^2} \left\|x_t^{ij} - F_{ij}\right\|^2 \mathrm{d}x_t^{ij}. \tag{13}$$

Although we do not know the distribution of real gradients, this means value can be estimated by sampling. Let S be the set of user i such that $y_{ij} = 1$.

$$\int P_{\text{real}}\left(x_t^{ij}\right) \frac{1}{2\sigma^2} \left\|x_t^{ij} - F_{ij}\right\|^2 \mathrm{d}x_t^{ij} \sim \sum_{i \in S} \frac{1}{2\sigma^2} \left\|x_t^{ij} - F_{ij}\right\|^2. \tag{14}$$

And in our algorithm, for a given item j, all the users will use the same F—in other words, we F_{ij} is independent of i. then the above equation is a function of the quadratic form.

$$\sum_{i \in S} \frac{1}{2\sigma^2} \left\|x_t^{ij} - F_j\right\|^2. \tag{15}$$

In order to minimize this K-L distance, we should set $F_j = \sum_{i\in S}(x^{ij}/\#S)$. $G_j(t) \approx (1/m)\sum_{i\in S} u_i(t)(r_{ij} - u_i^T(t)v_j(t)) = \nabla_{v_j} L(v_j(t))$. However, at time t, the user i can not get the current gradient $\nabla_{v_j} L(v_j(t))$. However, in the following section, we will show that in our algorithm we can estimate it from the previous gradient $\sum_i g_{ij}(t-1)$.

4.2. Algorithmic Description. In Algorithm 1 with Gaussian random mechanism, we can see that the central server will receive the gradients submitted from the users, whose summation is as follows:

$$G_j = \sum_{i \text{ with } y_{ij}=1} u_i\left(r_{ij} - u_i^T v_j\right) + N\left(0, \sigma^2\right) + \sum_{i \text{ with } y_{ij}=0} F_j + N\left(0, \sigma^2\right). \tag{16}$$

Suppose $F_j = 0$, G_j is just a Langevin stochastic gradient [30] whose expected value is the total gradient. When $F_j \neq 0$, using G_j to update the parameters will generally influence the accuracy of the model. One way to solve this problem is to subtract a value in the central server.

$$\begin{aligned} G_j = &\sum_{i \text{ with } y_{ij}=1} u_i\left(r_{ij} - u_i^T v_j\right) + N\left(0, \sigma^2\right) \\ &+ \sum_{i \text{ with } y_{ij}=0} F_j + N\left(0, \sigma^2\right) - N_j F_j. \end{aligned} \tag{17}$$

In order to determine the value of N_j to make the F_j part small, we can use the Random Response mechanism.

The random Response mechanism [31] is a well-known method to obtain statistical information on sensitive issues, e.g., the proportion of AIDS. In our algorithm, we will use the Random Response mechanism to count the number of $y_{ij} = 0$ items, which is used for the central server to correct the sum of the gradients.

The procedure of the Random Response mechanism is that the responder will give the true answer with probability $p > 0.5$, and with probability 1-p, the answer will give an opposite answer.

Theorem 4 (Warner, 1965, in [31]). Suppose the number of the answer of $y = 0$ is n_1, and the total number of the responders is n. If $p \neq (1/2)$, $(p - 1/2p - 1) + (n_1/(2p-1)n)$ is an unbiased estimate of the $y = 0$ ratio with variance $(\theta(1-\theta)/n) + (p(1-p)/n(2p-1)^2)$, where θ is the real ratio of $y = 0$ items.

The variance is $O(1/n)$, so if the total number of the users is large enough, with a high probability, $\widehat{\theta} \approx \theta$.

The whole process is shown in Algorithm 2.

It is easy to see that, in the central server, the update process has the following forms:

$$v_j(t+1) = v_j(t) - \eta\left(\nabla_{v_j} L\left(v_j(t), z\right) + \left(\frac{\text{num}_j - \theta_j n_j}{n}\right)\Delta V_j\right), \tag{18}$$

where $\nabla_{v_j} L(v_j(t), z)$ is the sampling stochastic gradient and num_j is the number of $y_{ij} = 0$ terms in sampling.

As for ΔV_j, we know that

$$\begin{aligned} \Delta V_j(t) &= \frac{\sum_{i\in S} g_{ij}(t-1) + \left(\#\{i \notin S\} - \theta_j n_j\right)\Delta V_j(t-1)}{\left(1-\theta_j\right)n_j} \\ &\sim \frac{\sum_{i\in S} g_{ij}(t-1)}{\#S}. \end{aligned} \tag{19}$$

Note that since the regularization term bound the norm of matrix U and V, there exits a small constant β to make the loss function L(u, v) to be $\beta-$ smooth, that is to say,

$$\left\|\nabla_{v_j} L\left(v_j(t-1)\right) - \nabla_{v_j} L\left(v_j(t)\right)\right\|_2 \leq \beta\left\|v_j(t-1) - v_j(t)\right\|_2. \tag{20}$$

Since $(1/\eta) \gg \beta$, $\nabla_{v_j} L(v_j(t-1))$ is a good approximation of $\nabla_{v_j} L(v_j(t))$.

So we have the folowing:

$$v_j(t+1) = v_j(t) - \eta\left(\nabla_{v_j} L\left(v_j(t), z\right) + \mu\nabla_{v_j} L\left(v_j(t-1), z\right)\right) + \zeta. \tag{21}$$

One can easily prove that the variances of all these estimations are $O(1/n)$.

4.3. The Influence of Model Accuracy. We can see the form of updating rule (21) is a stochastic gradient descent with time delay. It can be shown that even if μ a not small, time delay SGD will still have good convergence.

The convergence of SGD with time delay is proved in [32]. In this paper, Lian proved the convergence of asynchronous stochastic gradient descent which has the same form as equation (21).

Theorem 5 (Theorem 1 in [32]). Assume the loss function is $\beta-$ smooth, η is the learning rate, B is the batch size, and T is the time delay. If

$$\beta B\eta + 2\beta^2 B^2 T\eta \sum_k^T \eta \leq 1, \tag{22}$$

after K iterations, we have with high probability,

$$\min_k \left\|\nabla f\left(x_k\right)\right\|^2\Big) \leq 4\sqrt{\frac{\left(f\left(x_1\right) - f(x*)\right)\beta}{BK}}\sigma, \tag{23}$$

Where $f(x*)$ is the global minimum of f and σ is the standard deviation of stochastic gradients.

Proof of Theorem 5. In this case, the stochastic gradients $G_{m,t}$ sent by the node m at time t can be written as $G_{m,t} = \nabla f(x_{t-\tau_{m,t}}) + \zeta_{t,m}$, where $\tau_{m,t}$ is the time delay of the gradient and $\zeta_{t,m}$ is the noise (including noise from the stochastic gradients and the Gaussian noise we added). In our case, $\zeta_{t,m}$ is a sub-Gaussian random variable. To simplify the description, we assume $\zeta_{t,m}$ is σ-sub-Gaussian.

$$
\begin{aligned}
f(x_{\tau+1}) - f(x_0) &= \sum_{k=0}^{\tau} f(x_{k+1}) - f(x_k) \\
&\le \sum_{k=0}^{\tau} -\frac{B\eta}{2}\left\|\nabla f(x_k)\right\|^2 + \left(\frac{3\eta^2 L}{4} - \frac{\eta}{2B}\right)\left\|\sum_{m=1}^{B}\nabla f\left(x_{k-\tau_{k,m}}\right)\right\|^2 \\
&\quad + \beta^2 TB\eta \sum_{j=k-T}^{k-1}\eta^2\left\|\sum_{m=1}^{B}\nabla f\left(x_{j-\tau_{j,m}}\right)\right\|^2 - \eta\langle\nabla f(x_k), \sum_{m=1}^{B}\zeta_{k,m}\rangle + \frac{3\eta^2\beta}{2}\left\|\sum_{m=1}^{B}\zeta_{k,m}\right\|^2 \\
&\quad + \beta^2 B\eta^3\left\|\sum_{j=k-\tau_k^{\max}}^{k-1}\sum_{m=1}^{B}\zeta_{j,m}\right\|^2 = \sum_{k=0}^{\tau} -\frac{B\eta}{2}\left\|\nabla f(x_k)\right\|^2 \\
&\quad + \sum_{k=0}^{\tau}\left(\frac{3\eta^2\beta}{4} - \frac{\eta}{2B}\right)\left\|\sum_{m=1}^{B}\nabla f\left(x_{k-\tau_{k,m}}\right)\right\|^2 + \beta^2 TB\eta\sum_{j=k-T}^{k-1}\eta^2\left\|\sum_{m=1}^{B}\nabla f\left(x_{j-\tau_{j,m}}\right)\right\|^2 \\
&\quad - \sum_{k=0}^{\tau}\eta\langle\nabla f(x_k), \sum_{m=1}^{B}\zeta_{k,m}\rangle + \frac{3\eta^2\beta}{2}\left\|\sum_{m=1}^{B}\zeta_{k,m}\right\|^2 + \beta^2 B\eta^3\left\|\sum_{j=k-\tau_k^{\max}}^{k-1}\sum_{m=1}^{B}\zeta_{j,m}\right\|^2 \\
&\le \sum_{k=0}^{\tau} -\frac{B\eta}{2}\left\|\nabla f(x_k)\right\|^2 + \sum_{k=0}^{\tau}\left(\eta^2\left(\frac{3\beta}{4} + \beta^2 BT^2\eta\right) - \frac{\eta}{2B}\right)\left\|\sum_{m=1}^{B}\nabla f\left(x_{k-\tau_{k,m}}\right)\right\|^2 \\
&\quad \underbrace{-\sum_{k=0}^{\tau}\eta\langle\nabla f(x_k), \sum_{m=1}^{B}\zeta_{k,m}\rangle + \sum_{k=0}^{\tau}\frac{3\eta^2\beta}{2}\left\|\sum_{m=1}^{B}\zeta_{k,m}\right\|^2}_{T_{2,a}} + \underbrace{\sum_{k=0}^{\tau}\beta^2 B\eta^3\left\|\sum_{j=k-\tau_k^{\max}}^{k-1}\sum_{m=1}^{B}\zeta_{j,m}\right\|^2}_{T_{2,b}}.
\end{aligned}
\tag{24}
$$

In order to estimate $T_2 = T_{2,a} + T_{2,b}$, we can use lemmas in [33].

Let $\zeta_k = (1/B)\sum_{m=1}^{B}\zeta_{k,m}$. With probability $1 - e^{-\iota}$, we have the following:

$$
-\sum_{k=0}^{\tau}\eta\langle B\nabla f(x_k), \zeta_k\rangle \le \frac{\eta B}{8}\sum_{k=0}^{\tau}\left\|\nabla f(x_k)\right\|^2 + c\eta\sigma^2\iota. \tag{25}
$$

This is from Lemma 30 in [33].

With high probability,

$$
\sum_{k=0}^{\tau}\frac{3\eta^2\beta}{2}\left\|\sum_{m=1}^{B}\zeta_{k,m}\right\|^2 \le \frac{3\eta^2\beta}{2}Bc\sigma^2(\tau + 1 + \iota). \tag{26}
$$

And with high probability,

$$
\begin{aligned}
&\sum_{k=0}^{\tau}\beta^2 B\eta^3\left\|\sum_{j=k-\tau_k^{\max}}^{k-1}\sum_{m=1}^{B}\zeta_{j,m}\right\|^2 \\
&\le \beta^2 TB\eta^3 Bc\sigma^2\left(\frac{\tau}{TB\eta + 1 + T + \iota}\right) \\
&\le \frac{\eta^2 L}{2}Bc\sigma^2(\tau + 1 + \iota).
\end{aligned}
\tag{27}
$$

We have the following:

$$
T_1 \le \frac{\eta B}{8}\sum_{k=0}^{\tau}\left\|\nabla f(x_k)\right\|^2 + c\eta\sigma^2\iota + 2\eta^2\beta Bc\sigma^2(\tau + 1 + \iota), \tag{28}
$$

$\eta^2((3\beta/4) - \beta^2 MT^2\eta) - (\eta/2B) < 0$. With probability at least $1 - 3e^{-\iota}$,

$$
\begin{aligned}
f(x_{\tau+1}) - f(x_0) &\le \sum_{k=0}^{\tau} -\frac{3B\eta}{8}\left\|\nabla f(x_k)\right\|^2 + c\eta\sigma^2\iota \\
&\quad + 2\eta^2\beta Bc\sigma^2(\tau + 1 + \iota).
\end{aligned}
\tag{29}
$$

The theorem follows.

This theorem has the same form as the convergence theorem of general and SGD, and in our case, we have $T = 1$. So we can show this time delay will not influence the convergence. □

4.4. Privacy Loss in the Random Response Mechanism. At the start of our algorithm, we need to use the Random response mechanism to estimate the ratio of y_{ij}, which will cause a privacy loss. However, we can show that since we need a large number of iterations in the machine learning algorithm, the initial privacy loss is insignificant.

Input: Redefined iteration number k, learning rate η, probability p for Random Response and Standard deviation of Gaussian distribution σ.
Output: Item profile matrix V
For all items j, use the probability p Random Response method to estimate the ratio of the users with $y_{ij} = 0$ as θ_j.
Randomly initialize $u_i(0), v_j(0)$ for all i and j.
for $t = 1, 2, 3 \ldots$ **do**
  Initialize $G_j = 0$, $n_j = 0$ for all $j = 1, 2, \ldots,$ n in central server.
  for $i = 1, 2, 3, \ldots, m$ **do**
    On user i: sample B items $S = \{S_1, S_2, \ldots, S_B\}$ uniformly from$\{1, 2, 3, \ldots, n\}$
    for $j \, in \, S$ **do**
      $n_j = n_j + 1$
      if $y_{ij} = 1$ **then**
        $g_{ij} = u_i(r_{ij} - u_i^T v_j)$
        Draw $g'_{ij} \sim N(x_{ij}, \sigma^2)$
      end
      else
        if $t \neq 1$ **then**
          $F_{ij} = \Delta V_j$
        end
        else
          $F_{ij} = u_i(0 - u_i^T v_j)$
        end
        $g_{ij} = F_{ij}$
        Draw $g'_{ij} \sim N(x_{ij}, \sigma^2)$
      end
    end
    $G_j = G_j + x'_{ij}$.
  end
  for $j = 1, 2, \ldots, n$ **do**
    if $n_j = 0$ **then**
    $G_j = 0$
    $\Delta V_j = 0$
    end
    else
      $G_j = G_j - \theta_j n_j \times \Delta V_j(t-1)$
      $\Delta V_j(t) = (G_j / ((1 - \theta_j) n_j))$
      $G_j = (G_j / m)$
      $v_j = v_j + \eta G_j$
    end
  end
  for $i = 1, 2, 3, \ldots, m$ **do**
    Update u_i on the local device by gradient descent.
  end
end

Algorithm 2: Noisy matrix factorization with fake gradient.

It is easy to prove that the Random response mechanism is $\ln(p/1-p)$-Differential Privacy. We know from Theorem 3 that $\varepsilon \sim O(\sqrt{k \ln(1/\delta)})$ after k iterations. If n is large enough, we can choose a p near 0.5, and when k is large, $\ln(p/1-p)$ will be much less then ε.

Noting that the K-L distance for a length k sequence is O(k), the discussion on the K-L distance is the same.

5. Experiments

We now show the performance of our algorithm. We evaluate three types of privacy gradient descent algorithms:

(i) Algorithm 1, the noisy gradient descent with $F_{ij} = u_i(0 - u_i^T v_j)$. The users will submit a gradient $F_{ij} = u_i(0 - u_i^T v_j) + \zeta$ if $y_{ij} = 0$, where ζ is a $N(0, \sigma^2)$ Gaussian random variable.

(ii) Algorithm 2, noisy gradient descent with $F_{ij} = \zeta$.

(iii) Algorithm 3, our algorithm in this paper.

In the $F_{ij} = 0$ case, the only noise in the total gradient is caused by Gaussian noise added to the users' device. This algorithm will be accurate but has no ability to protect the item's privacy. We will show that the performance of our algorithm is very close to the case $F_{ij} = 0$ and much better than the algorithm using fake ratings.

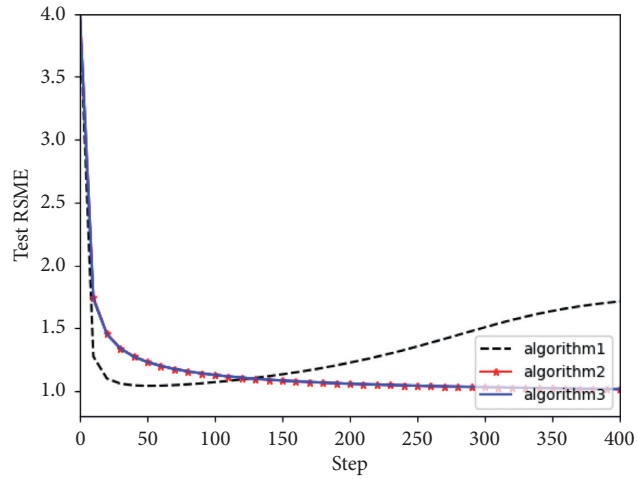


Figure 1: RMSE with 50% density.

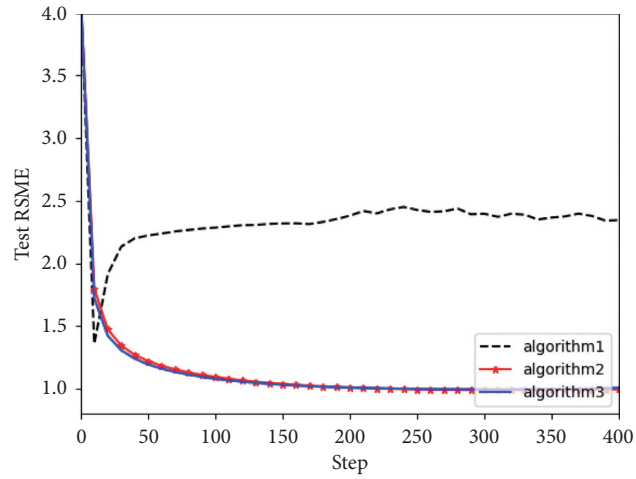


Figure 2: RMSE with 75% density.

We test on MovieLens 100k dataset [34]. This version contains 100k ratings of 1682 movies submitted by 982 users. This dataset is very sparse. In order to test the performance in different situations of sparsity, for every user, we choose a set F of items to be selected to provide fake gradients. We consider different cases that $\#F = \#S$ (50% fake gradient density), $\#F = 3\#S$ (75% fake gradient density) to test the algorithm. We set the profile vector dimension $d = 15$, regularization parameters $\lambda = 0.001$, learning rate $\eta = 0.1$, $\sigma^2 = 1$ and use AdaDelta to optimize. The test RMSE is shown in Figures 1 and 2.

After 400 iterations, the test RMSE is listed in Table 2.

We see that when the density of fake rating increases, the test RSME of fake rating algorithm is growing rapidly, and the performance algorithm is very close to the zero mean fake gradient algorithm.

6. Related Work

Differential privacy introduced by Dwork [13] is a very strong guarantee to protect privacy. The original version of differential privacy consider a trusted server to provide data to queriers, and the aim is to prevent access to user privacy from queries.

Local differential privacy algorithm, such as RAPPORT [22], is to make sure the central server can not access the privacy of the users. The main technology is to add some noise before submitting the data to the server. In the Chrome browser, Google uses a randomized response mechanism to collect the data of the users' clicks. Also, there are many works to use local differential privacy to perform machine learning algorithms. For example, Google uses local differential privacy Federated Learning [35] to learn a language model in order to improve the performance of the inputting method.

One of the difficulties in differential privacy machine learning is that when training a model using many iterations, the privacy guarantees will decline rapidly. Differential privacy for multi-iterations is studied in [25, 26] and a much tighter composition theorem is given.

Private recommender system is studied by many authors such as [17–20, 36, 37]. References [17, 18] are based on a matrix factorization recommender system. The algorithm is to adding some noise in users' devices locally to protect

Table 2: RMSE in experiments.

Density (%)	Method RMSE		
	$F_{ij} = u_i(0 - u_i^T v_j)$	$F_{ij} = 0$	Our algorithm
50	1.7141140	1.0161815	1.0167035
75	2.3660726	0.9966793	1.0073330

privacy. The algorithm in [17] can protect both the ratings and the items of the user. Their work is based on the work in [24], where they propose a new randomization mechanism and show that their mechanism is better when the dimension of data is large.

7. Conclusion

In this paper, we propose a novel privacy matrix factorization algorithm. In our algorithm, we use the Random Response method to estimate the selection ratios of the items, and then we use the average value of the gradients in the previous time as the fake gradient to be sent to the central server. Using our method, we can improve the indistinguishability of the real gradient and fake distributions so that improve the ability to protect user private items. Meanwhile, we show that our algorithms will not cut down the accuracy of the model since the updating rule can be reduced to SGD with time delay, which can be proved to convergence to gradient zero points.

Data Availability

The Movielens-100K, http://files.grouplens.org/datasets/movielens/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by the Fundamental Research Funds for the Central Universities (grant number: 2020JBM002) and the National Key Research and Development Program of China (grant no. 2018YFC0831703).

References

[1] Z. Zhang, Y. Liu, Z. Zhang, and B. Shen, "Fused matrix factorization with multi-tag, social and geographical influences for poi recommendation," *World Wide Web*, vol. 2222 pages, 2018.

[2] Z. Zhang, Y. Liu, G. Xu, and H. Chen, "A weighted adaptation method on learning user preference profile," *Knowledge-Based Systems*, vol. 112, no. 15, pp. 114–126, 2016.

[3] T. G. Alexandru and C. Pupăză, "Machine learning generalization of lumped parameter models for the optimal cooling of embedded systems," *Studies in Informatics and Control*, vol. 29, no. 2, pp. 169–177, 2020.

[4] I. Stoica, "Solving system problems with machine learning," *Studies in Informatics and Control*, vol. 28, no. 2, pp. 119–132, 2019.

[5] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, pp. 77–92, 2019.

[6] V. Oona, "Using data mining methods to solve classification problems in financial-banking institutions," *Economic Computation and Economic Cybernetics Studies and Research/Academy of Economic Studies*, vol. 54, no. 1, pp. 159–176, 2020.

[7] K. J. Ban, "Implementation of artificial intelligence system and traditional system: a comparative study," *Journal of System and Management Sciences*, vol. 66, 2019.

[8] S. Y. Y. Ji and L. J. Ku, "A blockchain and internet of things based architecture design for energy transaction," *Journal of System and Management Sciences*, vol. 63, 2020.

[9] S. H. L. Z. X. Li and J. Pan, "A machine learning based method for customer behavior prediction," *Tehnicki vjesnik-Technical Gazette*, vol. 72, 2019.

[10] N. Z. D. Qin and L. L. Yu, "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video," *Tehnicki vjesnik-Technical Gazette*, vol. 58, no. 3, 2018.

[11] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, "You might also like: privacy risks of collaborative filtering," *Security & Privacy*, vol. 39, no. 5, 2012.

[12] F. Dan and J. Riedl, "Do you trust your recommendations? an exploration of security and privacy issues in recommender systems," in *Proceedings of the International Conference On Emerging Trends In Information & Communication Security*, Freiburg, Germany, June 2006.

[13] C. Dwork, K. Kenthapadi, F. Mcsherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Proceedings of the International Conference On Advances In Cryptology-Eurocrypt*, Saint Petersburg, Russia, June 2006.

[14] Z. Liu, Y. X. Wang, and A. Smola, "Fast differentially private matrix factorization," *Machine Learning*, vol. 9, 2015.

[15] T. Zhu, L. Gang, Y. Ren, W. Zhou, and X. Ping, "Differential privacy for neighborhood-based collaborative filtering," in *Proceedings of the IEEE/ACM International Conference On Advances In Social Networks Analysis & Mining*, Ontario Canada, August 2013.

[16] A. Machanavajjhala, A. Korolova, and A. D. Sarma, "Personalized social recommendations: accurate or private," *Proceedings of the Vldb Endowment*, vol. 4, no. 7, 2011.

[17] H. Shin, S. Kim, J. Shin, and X. Xiao, "Privacy enhanced matrix factorization for recommendation with local differential privacy," *IEEE Transactions on Knowledge & Data Engineering*, vol. 99, p. 1, 2018.

[18] A. Berlioz, A. Friedman, M. A. Kaafar, R. Boreli, and S. Berkovsky, "Applying differential privacy to matrix factorization," in *Proceedings of the Acm Conference On Recommender Systems*, Vienna, Austria, September 2015.

[19] J. Hua, X. Chang, and Z. Sheng, "Differentially private matrix factorization," in *Proceedings of the International Conference on Artificial Intelligence*, Deigo, CL, USA, June 2015.

[20] Y. Shen and H. Jin, "Privacy-preserving personalized recommendation: an instance-based approach via differential privacy," in *Proceedings of the IEEE International*

Conference on Data Mining, Shenzhen, China, December 2014.

[21] Y. Shen and H. Jin, "Epicrec: Towards practical differentially private framework for personalized recommendation," in *Proceedings of the Acm Sigsac Conference on Computer & Communications Security*, London, UK, November 2016.

[22] U. Erlingsson, V. Pihur, and A. Korolova, "Rappor:randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the Acm Sigsac Conference on Computer & Communications Security*, Scottsdale, Ariz, USA, November 2014.

[23] Apple's Differential Privacy Collecting Data," 2016, https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/.

[24] T. T. Nguyen, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," *Computer Science Databases*, vol. 16, 2016.

[25] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," *Foundations of Computer Science Annual Symposium on*, vol. 26, no. 2, pp. 51–60, 2010.

[26] M. Abadi, A. Chu, I. Goodfellow et al., "Deep learning with differential privacy," in *Proceedings of the Acm Sigsac Conference On Computer & Communications Security*, Vienna, Austria, October 2016.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2003.

[28] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," *Computer and Communications Security*, vol. 28, pp. 43–54, 2016.

[29] C. Dwork, "Calibrating noise to sensitivity in private data analysis," *Lecture Notes in Computer Ence*, vol. 3876, no. 8, pp. 265–284, 2012.

[30] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient langevin dynamics," *Proceedings of Machine Learning Research*, vol. 65, pp. 1–43, 2017.

[31] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[32] X. Lian, Y. Huang, Y. Li, and L. Ji, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Proceedings of the International Conference On Neural Information Processing Systems*, Montreal, Canada, December 2015.

[33] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, "Stochastic gradient descent escapes saddle points efficiently," *arXiv: Learning*, vol. 31, 2019.

[34] F. M. Harper and J. A. Konstan, "The movielens datasets: history and context," *Ksii Transactions on Internet and Information Systems*, vol. 5, no. 4, p. 19, 2016.

[35] R. C. Geyer, T. Klein, and M. Nabi, *Differentially private federated learning: a client level perspective*, 2017.

[36] H. Kikuchi and A. Mochizuki, "Privacy-preserving collaborative filtering using randomized response," *Journal of Information Processing*, vol. 21, no. 4, pp. 671–676, 2012.

[37] Y. Xin and T. Jaakkola, "Controlling privacy in recommender systems," in *Proceedings of the International Conference On Neural Information Processing Systems*, Long Beach, CA, USA, December 2014.

Hindawi
Complexity
Volume 2020, Article ID 6689304, 9 pages
https://doi.org/10.1155/2020/6689304

Research Article

Adaptive Attention with Consumer Sentinel for Movie Box Office Prediction

Kaicheng Feng and Xiaobing Liu

School of Economics and Management, Dalian University of Technology, Dalian 116024, China

Correspondence should be addressed to Kaicheng Feng; 11311062@mail.dlut.edu.cn

Received 4 November 2020; Revised 14 November 2020; Accepted 19 November 2020; Published 7 December 2020

Academic Editor: Abd E. I.-Baset Hassanien

To improve the movie box office prediction accuracy, this paper proposes an adaptive attention with consumer sentinel (LSTM-AACS) for movie box office prediction. First, the influencing factors of the movie box office are analyzed. Tackling the problem of ignoring consumer groups in existing prediction models, we add consumer features and then quantitatively analyze and normalize the box office influence factors. Second, we establish an LSTM (Long Short-Term Memory) box office prediction model and inject the attention mechanism to construct an adaptive attention with consumer sentinel for movie box office prediction. Finally, 10,398 pieces of movie box office dataset are used in the Kaggle competition to compare the prediction results with the LSTM-AACS model, LSTM-Attention model, and LSTM model. The results show that the relative error of LSTM-AACS prediction is 6.58%, which is lower than other models used in the experiment.

1. Introduction

The movie box office, as an indicator of the level of film development, has attracted great attention from all walks of life. At present, the prediction of the movie box office has become one of the hottest research by scholars [1]. Linear regression and nonlinear regression models used to construct social media-driven movie box office prediction models were proposed [2]. A new method of movie box office prediction based on two-level and twice proxy variables was proposed [3], which can predict the first week's box office by using some preindicators obtained before the movie is released. A single influencing factor of the movie box office was mainly analyzed [4]. They analyzed the influence of celebrity effect on box office. They concluded that celebrity influence is positively related to box office. The competition factors with similar movie release time on the standard regression framework were tested, and a more simplified empirical model was proposed [5]. A BRP feedback neural network was proposed to solve movie box office prediction and classification problems [6]. The prediction model using the BP neural network has the following shortcomings. (1) Binary is used in the discretization of the model to quantify the various influencing factors of the movie box office [7]. These variables are not processed according to the actual situation, and the differences between different influencing factors cannot be fully expressed. (2)When using BP neural network for training, it is easy to fall into the problem of local minimization [8].

LSTM [9] is a time recurrent neural network. A movie box office prediction based on the LSTM model was proposed [10]. This model can solve the BP neural network using only simple Boolean coefficient values. It can map as many movie box office influence factors as possible in the input and output. However, its analysis of film sale influencing factors is not comprehensive, and the predicted results still have large relative errors.

Tackling the current movie box office prediction problem, this paper proposes an adaptive attention LSTM model with consumer sentinel. Compared with traditional LSTM, this model proposes an attention with consumer sentinel. On the one hand, it can fully consider the impact of movie consumer information on the movie box office and improve the model input. On the other hand, adaptive attention can

capture the input affective information more vigorously, thereby further improving the prediction accuracy. Specifically, the model is based on the LSTM model injecting the adaptive attention (AAM) with consumer sentinel. Consumer sentinel can identify the influence of the influencing factors of the movie box office from more dimensions and solve the long-standing problem of ignoring consumer information in box office forecasting. The use of LSTM takes into account the random volatility and long-time span of the movie box office. LSTM also remembers the information for a long time to predict the movie box office. Injecting an adaptive attention can capture affective input information, which provides a guarantee for the accuracy of the movie box office prediction results. The proposed model provides a certain reference value for film investors in film risk control, and it can have a certain planning value for film release schedule and has practical application prospects. The contribution of this paper can be summarized as follows. (1) To improve the movie box office prediction accuracy, this paper proposes an LSTM model with an AAM and consumer sentinel (LSTM-AACS). It can better capture consumer characteristics, thereby improving prediction accuracy. (2) The LSTM-AACS model is applied to the prediction of the movie box office and achieves good results. The results show that the relative error of LSTM-AACS prediction is 6.58%, which is lower than other models used in the experiment.

2. Literature Review

There are many factors including investment, director, actors, and sequel and play a role in promoting and guiding the film's box office. In [11], six independent variables of film investment were selected; film quality, director, actors, film sequel, and piracy. They established a linear regression model for influencing factors and movie box office. A semiparametric method was proposed to deal with random effects in a nonparametric way [12]. The example of comparing the reviews of movie critics uses the logit model of the adjacent category and the logit model of the related baseline category. Although this method eliminates the influence of extreme data, it also makes the use of data information insufficient.

The above research provides an important reference when selecting factors affecting the movie box office. Then, they used the Sawhney and Eliashberg model to predict the cumulative number of viewers of the movie after a few weeks of its release [13]. Its practical significance is that, during the life cycle of movie release, movie theaters can dynamically adjust the projection strategy. For example, movie producers can expand or reduce the number of theaters showing the movie, change the projection period, and so on. However, this method has the following shortcomings. (1) When using the multiple linear regression algorithm to predict the cumulative audience in the first week, few film influencing factors (number of film copies, user ratings, number of theaters, and audience age) are considered, and it did not consider the special attributes of the movie to attract the audience. This led to an excessively large prediction error for the first week. (2)This kind of error will accumulate when the diffusion model is used to predict the number of viewers in the next few weeks, which will affect the final prediction accuracy.

Based on the multilayer neural network algorithm, multiple movie attributes that affect the box office were combined [14]. They proposed a movie box office classification model and used the classification accuracy as the main index to evaluate the classification performance of the model. They achieved good classification results. However, this method uses binary discrete numbers to quantify the various influencing factors of the movie box office, which is obviously a vague processing method. These variables are not quantified according to the actual situation, so they cannot fully reflect the different variables in the influencing factors. In addition, the classification of the movie box office in the output layer of the prediction model is also vague, making the classification of each movie box office level too large. Such classification is of little relevance value for film investors and movie theaters to control the cost of film production and screening. A multimodal deep neural network for movie box office revenues prediction was proposed [15]. A CNN was built for extracting features from movie posters. Then, a multimodal deep neural network was built to leverage both movie poster features and other movie-related data for movie box-office revenues prediction. In addition, the features of CNN learned from movie posters were analyzed. However, the research did not focus on building more multimode DNNs, nor did it merge audio and video data related to movies. In [16], a hybrid social recommender system utilizing a deep autoencoder network is introduced. The proposed approach employs collaborative and content-based filtering, as well as users' social influence. The social influence of each user is calculated based on his/her social characteristics and behaviors on Twitter. For the evaluation purpose, the required datasets have been collected from MovieTweetings and Open Movie Database. However, the dataset used in this study is not comprehensive enough and may have limitations in prediction accuracy.

The LSTM-AACS model used in this paper uses the LSTM model based on the adaptive attention. A lot of work has been proposed for the LSTM model based on the attention. The attention-based LSTM model was proposed for financial time series prediction [15], and the model prediction can be intuitively understood through the attention vector. In addition, their focus on time and factors makes it easy for people to understand why certain trends are predicted when accessing a given time series table. They also modified the loss function of the attention model using weighted classification crossentropy. However, there is a shortcoming that the error is small in the long-term forecast, and the performance in the short-term forecast is not ideal, with high errors. A forecasting framework was established to predict the opening prices of stocks [16]. They processed stock data through a wavelet transform and used an attention-based LSTM neural network to predict the stock opening price, with excellent results. However, simply considering the impact of historical data on price trends is too singular and may not be able to fully and accurately forecast the price on a given day. An attention-based long

short-term memory network for aspect-level sentiment classification was proposed [17]. The attention mechanism can concentrate on different parts of a sentence when different aspects are taken as input. However, its flaw is that different aspects are input separately, and it does not realize modeling of multiple aspects simultaneously with the attention mechanism. An attention-based LSTM network is proposed for cross-language sentiment classification [18]. They use bilingual bidirectional LSTM to model the sequence of words in the source and target languages. Based on the particularity of sentiment classification tasks, they proposed a hierarchical attention model that was jointly trained with LSTM network. The model has achieved gratifying results on the benchmark dataset with Chinese as the source language and English as the target language. However, the problem is that the performance of the model is not evaluated on more datasets and more language pairs. An attention-based LSTM model for the task of hashtag recommendation was proposed [19]. They adopted the architecture of LSTM to avoid hand-crafted features. Their model incorporates topic modeling into the LSTM architecture through an attention mechanism and takes over the advantages of the both. Through evaluations run on a large dataset from Twitter, they have demonstrated that the proposed method outperforms competitive baseline methods effectively. However, the present work does not consider the use of other types of data in microblogs for hashtag recommendation [20].

The main problems above are as follows. (1) It performs well in short-term prediction, but the effect is not ideal in long-term prediction. (2) The input data of the model is not comprehensive, which leads to the prediction results only in a certain dataset to achieve high prediction accuracy. (3) The influencing factors of the results in the prediction problem are not considered comprehensively, such as ignoring user information and resulting in low prediction accuracy. Based on the above problems, we propose an AAM for movie box office prediction with consumer sentinel. With consumer sentinel, it can solve the problem of ignoring consumer groups in previous predictions. AAM can capture effective input information well. Finally, the LSTM model based on the above two algorithms is used to predict the movie box office and compare with other models. Experiments show that the prediction accuracy of the AAM for the movie box office prediction model with consumer sentinel is better than other models used in the experiment.

3. Adaptive Attention Mechanism with Consumer Sentinel

3.1. Framework Design. The framework is shown in Figure 1. It can be seen that this paper adds consumer information to the previous movie box office influencing factors and injects an AAM into the LSTM neural network (its structure is shown in the blue box, and consumer sentinel are input into the model as features and then combined with the attention mechanism to train the LSTM model). This improves the prediction accuracy.

3.2. Determination of Influencing Factors

3.2.1. Factors of a Movie. This paper uses the statistical analysis of the historical box movie office data in China combined with the actual situation of the movie market. The paper selects the director, actor, film genre, nation, and release data as the film's own influencing factors (as the film's information input). This paper then assigns different weights to each factor. The calculation method will be explained in detail in Section 4.1.

3.2.2. Consumer Groups. Based on the consideration of a movie's own influencing factors mentioned in Section 3.2.1, this paper adds the age information of movie consumer groups. This is because every movie must have its audience. For example, military subjects are more suitable for viewing by teenagers and above, while cartoons have more children as the audience. Generally, elderly people rarely go to the cinema to watch movies and so on. The age information of consumers is used as input information, and weights are assigned to jointly predict the final box office of the movie.

3.3. Long- and Short-Term Memory Network Layer. LSTM is an improved RNN (Recurrent Neural Network) model that solves the problems of gradient explosion or gradient disappearance during RNN training. Different from the single tanh loop structure in standard RNN, LSTM is a special network with three "gates" [21,22]. They are the forget gate, input gate, and output gate. The forget gate is responsible for choosing to forget invalid information in the past. The input gate is responsible for determining that useful new information is stored in the cell state. The output gate determines the output information. The process of the memory module for status update and information output is as follows:

(1) The core of LSTM is cell: cell state is the memory transmission belt of the entire module that changes over time. The conveyor belt itself cannot control which information is memorized. The forget gate, input gate, and output gate play a controlling role.

(2) Forget state information: select the input x_t at the current moment and the memory unit state information h_{t-1} at the previous moment, and then use the sigmoid function to output a value of [0, 1] to indicate the degree to which historical information needs to be retained:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \tag{1}$$

(3) Update the status information and store useful new information in the cell state. First, calculate the value of the input gate. The function of the input gate is to control how the current data input affects the state value of the memory unit. Then, calculate the candidate memory unit information $\widetilde{C}_t$ at the current time t, which contains the new information to be added. Finally, merge the old cell state $C_{t-1} * f_t$

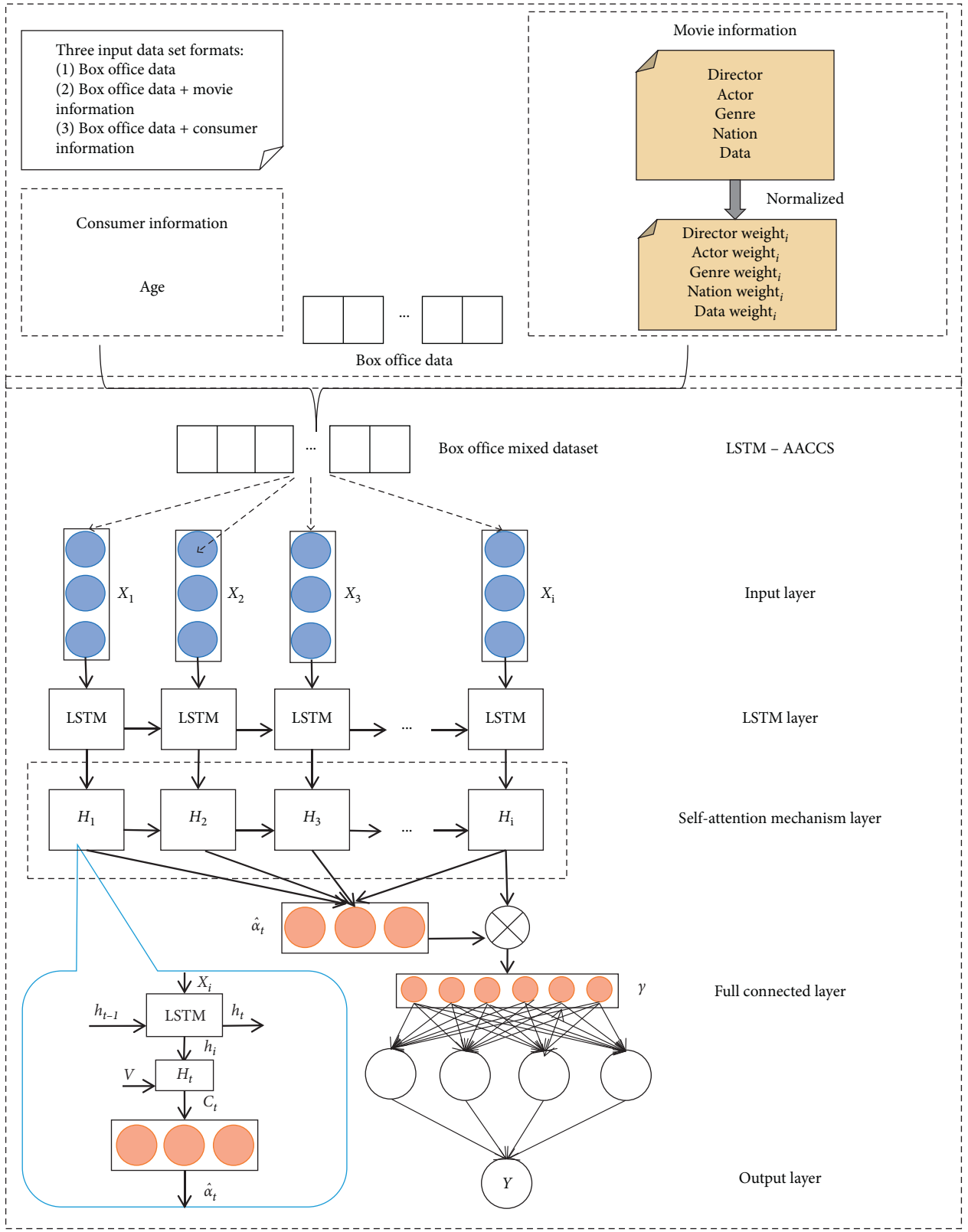


Figure 1: Overall framework of the model.

(used for forgetting) with the new candidate information $i_t * \widetilde{C}_t$ to determine the updated information:

$$\begin{aligned} i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \\ \widetilde{C}_t &= \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \\ C_t &= f_t * C_{t-1} + i_t * \widetilde{C}_t. \end{aligned} \tag{2}$$

(4) Output information: first determine which part of the state will be outputted. Finally, obtain the memory unit output information at the current time after the value of the output gate and the state information of the memory unit undergo tanh transformation:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \tag{3}$$

$$h_t = o_t * \tanh\left(C_t\right). \tag{4}$$

3.4. Adaptive Attention Mechanism Layer. This paper adds an AAM layer [22] to the method, which can better capture the affective information in the movie box office data and grasp the core data information. It overcomes the problem of the standard LSTM model using the same state vector in each step of the prediction, which results in the inability to fully learn the detailed information of the sequence encoding during the prediction. The specific extension method is based on the original LSTM model, adding two formulas:

$$\begin{aligned} g_t &= \sigma\left(W_i x_t + W_h h_{t-1}\right), \\ s_t &= g_t \odot \tanh\left(C_t\right), \end{aligned} \tag{5}$$

where W_i and x_t are the input of LSTM and $W_h \in R^{d \times d}$ is the parameter matrix that the subsequent model needs to train. C_t is a memory cell, called sentinel gate. It is similar to the input gate, forget gate, and output gate in LSTM. The structure of the formula is similar to (4). The vector $\hat{c}$ in the AAM can be expressed as follows:

$$\hat{c} = \beta_t s_t + \left(1 - \beta_t\right) c_t, \tag{6}$$

where $\beta_t \in [0, 1]$ can be regarded as a sentinel gate in the true sense controlling the degree to which the model pays attention to C_t. The representation of β_t is as follows:

$$\beta_t = w_h^T \tanh\left(W_s s_t + \left(W_g h_t\right)\right). \tag{7}$$

At the same time, the attention distribution α_t of the K areas of the AAM has also been expanded to $\hat{\alpha}_t$. The method is to splice an element after z_t:

$$\hat{\alpha}_t = \text{soft}\max\left(\left[z_t; \beta_t\right]\right), \tag{8}$$

where $\hat{\alpha}_t$ has $k+1$ elements, and the expression of z_t is as follows:

$$z_t = w_h^T \tanh\left(W_v V + \left(W_g h_t\right) 1^T\right). \tag{9}$$

Finally, the probability distribution of the matrix can be expressed as follows:

$$\gamma = \text{soft}\max\left(W_p\left(\hat{c} + t h_t\right)\right), \tag{10}$$

where W_p is the parameter matrix that the subsequent model needs to train. The output variable h_t finally passes through a fully connected layer and softmax classifier, which finalizes the prediction of movie box office.

4. Experiment and Discussion

4.1. Normalization of Impact Factors. This section will elaborate on the factors that affect the movie box office and give the corresponding definitions. At the same time, the quantification process of each attribute of the movie box office data will be given to prepare for the construction of the LSTM-AACS training set.

4.1.1. Director. Define movie box office influence of director i index as

$$\text{Dir}_i = \frac{\left(\sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk}\right)}{m}, \quad m = \min\{5, m\}, \tag{11}$$

where i means director number, j represents the jth movie filmed by director i, k indicates the week of the release, m means the m movies with the closest release time from the current time among all the movies filmed by director i, and b_{jk} represents the box office during the kth week of the latest jth movie. Furthermore, the box office influences weight DirectorWeight_i of the film directed by director i which can be obtained as follows:

$$\text{DirectorWeight}_i = \frac{\left(\lg\left(\text{Dir}_i / \text{Dir}_{\min}\right)\right)}{\left(\lg\left(\text{Dir}_{\max} / \text{Dir}_{\min}\right)\right),} \tag{12}$$

where

$$\begin{aligned} \text{Dir}_{\max} &= \max\{\text{Dir}_1, \text{Dir}_2, \ldots\}, \\ \text{Dir}_{\min} &= \min\{\text{Dir}_1, \text{Dir}_2, \ldots\}, \end{aligned} \tag{13}$$

where i represents the number of the director and Diri represents the influence of the ith director.

4.1.2. Actor. Define box office influence of actor i index as

$$A_i = \frac{\sum_{j=1}^{m} u_{ij}\left(\sum_{k=1}^{5} b_{jk}\right)}{m}, \quad m = \min\{5, m\}, \tag{14}$$

where i means actor number, j represents the jth movie filmed by the actor i, k indicates the week of the release, m means the m movies with the closest release time from the current time among all the movies filmed by actor i, b_{jk} represents the box office during the kth week of the latest jth movie, and u_{ij} is the participation coefficient of the jth movie actor i recently participated in, and it is defined as follows:

$$u_{ij} = \begin{cases} 1 - \dfrac{(n-1)}{10}, & n \in [1, 5], \\ 0.5, & n \in (5, +\infty), \end{cases} \tag{15}$$

where n is a positive integer, indicating the order of the actor i in the jth movie. Furthermore, the box office influence weight ActorWeight$_i$ of the film directed by actor i can be obtained as follows:

$$\text{ActorWeight}_i = \frac{(\lg(A_i/A_{\min}))}{(\lg(A_{\max}/A_{\min})),} \tag{16}$$

where

$$\begin{aligned} A_{\max} &= \max\{A_1, A_2, \ldots\}, \\ A_{\min} &= \min\{A_1, A_2, \ldots\}, \end{aligned} \tag{17}$$

where i represents the number of the actor and A_i represents the influence of the ith actor.

4.1.3. Movie Genre. Define movie box office influence of movie genre i index as

$$G_i = \sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk}, \tag{18}$$

where i means genre number ($i = 1,2, \ldots, 9$), k indicates the week of the release, m represents the week of screening of genre i, j represents the jth movie belonging to genre i, and b_{jk} represents the box office of the jth movie with content genre i in the kth week of its release. The box office influence weight GenreWeight$_i$ of the film of genre i can be obtained as follows:

$$\text{GenreWeight}_i = \frac{(\lg(G_i/G_{\min}))}{(\lg(G_{\max}/G_{\min})),} \tag{19}$$

where

$$\begin{aligned} G_{\max} &= \max\{G_1, G_2, \ldots\}, \\ G_{\min} &= \min\{G_1, G_2, \ldots\}. \end{aligned} \tag{20}$$

4.1.4. Nation. Define box office influence of movie nation i index as

$$N_i = \sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk}, \tag{21}$$

where i means nation number($1 \le i \le 5$), the value of i from 1 to 5 corresponds to Europe, America, Japan, Korea, Hong Kong, and Taiwan, Mainland China, and other regions, k indicates the week of the release, m represents the total number of movies in the distribution nation i belonging to the area, j represents the jth movie in the distribution nation i, and b_{jk} represents the box office during the kth week of the release of the jth movie in the distribution nation i. The box office influence weight NationWeight$_i$ of the film directed by nation i can be obtained as follows:

$$\text{NationWeight}_i = \frac{N_i}{\sum_{j=1}^{5} N_j}, \tag{22}$$

where i represents the serial number of the issuance area, N_i represents the influence of the issuance area i, and N_j represents the influence weight of the issuance area j.

4.1.5. Data. Define box office influence of release data i index as

$$D_i = \sum_{j=1}^{m} \sum_{k=1}^{4} b_{jk}, \tag{23}$$

where i means data number($1 \le i \le 4$), the value of i from 1 to 4 corresponds to the Lunar New Year file, the 51st file, the summer file, and the eleventh file. k indicates the week of the release, m represents the total number of movies with the release date in schedule i, j represents the jth movie belonging to data i, and b_{jk} represents the box office data generated during the kth week of the release date of the jth movie with the release date in schedule i. The weight DataWeight$_i$ measures the box office influence of the type on the movie attributable to that type:

$$\text{DataWeight}_i = \frac{(\lg(D_i/D_{\min}))}{(\lg(D_{\max}/D_{\min})),} \tag{24}$$

where

$$\begin{aligned} D_{\max} &= \max\{D_1, D_2, \ldots\}, \\ D_{\min} &= \min\{D_1, D_2, \ldots\}, \end{aligned} \tag{25}$$

where i represents the serial number of the data and D_i represents the influence of the data i.

4.1.6. Consumer Group. This paper is divided into 4 age groups: under 18, 18–45, 46–69, and over 69. Define box office influence of movie nation i index as

$$N_i = \sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk}, \tag{26}$$

where i means age group number($1 \le i \le 4$) the value of i ranges from 1 to 4 corresponding to ages under 18 years old (excluding 18 years old), 18–45 years old, 46–69 years old, and over 69 years old (excluding 69 years old), k indicates the week of the release, m represents the total number of movies in age group i, j represents the jth movie in the distribution age i, and b_{jk} represents the box office during the kth week of the release of the jth movie in the distribution age i. The box office influence weight AgeWeight$_i$ of the film indexed by age group i can be obtained as follows:

$$\text{AgeWeight}_i = \frac{(\lg(\text{Age}_i/\text{Age}_{\min}))}{(\lg(\text{Age}_{\max}/\text{Age}_{\min})),} \tag{27}$$

where

$$\begin{aligned} \text{Age}_{\max} &= \max\{\text{Age}_1, \text{Age}_2, \ldots\}, \\ \text{Age}_{\min} &= \min\{\text{Age}_1, \text{Age}_2, \ldots\}, \end{aligned} \tag{28}$$

where i represents the serial number of the age group and Age$_i$ represents the influence of the age group i.

Table 1: Specific error table.

Method	Average relative error (%)
LSTM	28.54
LSTM-attention	11.45
LSTM-AACS	6.58

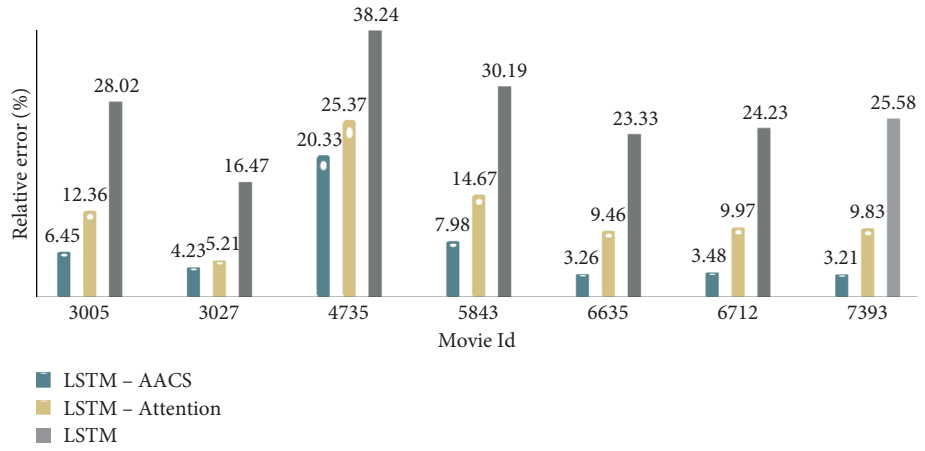


Figure 2: Relative error comparison.

4.2. Model Parameters. In the LSTM-AACS model, we set the dropout rate to 0.5. During training, a small batch stochastic gradient descent method is used to reduce the training loss, and the minibatch is set to 64. When analyzing the prediction results, the relative error is used for analysis. This paper uses the movie box office prediction dataset in the Kaggle competition. After obtaining the movie box office prediction data, the calculation formula for the relative error of the prediction result is as follows:

$$\text{Relative error} = \frac{\text{prediction} - \text{real}}{\text{real}} \times 100\%. \tag{29}$$

4.3. Comparison and Analysis of Experimental Results

4.3.1. Error Comparison. In the experiment, the results are analyzed by crossvalidation. This paper randomly takes out the first 3000 pieces of data from 10,398 pieces of data as the training set, and the last 7398 pieces of data as the test set. Learning is done 30 times when training the model, and then ten crossvalidation rounds are applied. Finally, the average relative error of these two models is shown in Table 1.

It can be seen from Table 1 that the average relative error of LSTM time series and LSTM-Attention, using ten crossvalidation, is higher than the relative error of the model proposed in this paper. It shows that the LSTM-AACS model is better than the LSTM model and the general LSTM model with attention for movie box office prediction.

Randomly selecting the prediction results of several movies from the test set, Figure 2 shows the comparison of partial relative errors of the two models under the training set.

From the seven randomly selected movies, it can be seen that the relative error predicted by the LSTM-AACS model in the test set is lower than the relative error predicted by the LSTM model and the LSTM-Attention model. The prediction results of the LSTM-AACS model is relatively more accurate, and the performance is improved.

4.3.2. Result Comparison. In order to make the model have both long-term and short-term prediction capabilities, we compare the long term prediction capabilities of the LSTM-AACS model, the LSTM-Attention model, and the LSTM time series model. We also compare and analyze the movie box office data in the Kaggle competition. Additionally, we choose the box office data of a Maoyan movie to predict the short term box office. Considering the classic movies of previous years, this paper chooses Dangal; My People, My Country; Wolf Warriors II and Fast & Furious 7 as the movies to predict their cumulative box office. Using these movies, this paper compares the actual value, predicted value, absolute difference, and relative error of the three models, respectively. The specific results are shown in Table 2.

As can be seen from Table 2, the relative error of the LSTM-AACS model in predicting the cumulative box office of the above four movies is lower than the relative error of the prediction results of the LSTM model and the LSTM-Attention model. This proves the feasibility of the LSTM-AACS model proposed in this paper in predicting movie box office. This also proves that the LSTM-AACS model can have a better evaluation effect on movie investors.

Table 2: Movie box office prediction results of previous classic movies.

Movie name	Real (100 million)	LSTM			LSTM-attention			LSTM-AACS		
		Prediction (100 million)	Absolute error (100 million)	Relative error (%)	Prediction (100 million)	Absolute error (100 million)	Relative error (%)	Prediction (100 million)	Absolute error (100 million)	Relative error (%)
Dangal	12.99	14.31	1.32	10.16	13.98	0.99	7.62	13.18	0.19	1.46
My People, My Country	31.71	35.42	3.71	11.70	33.67	1.96	6.18	32.59	0.88	2.78
Wolf Warriors II	56.92	60.21	3.29	5.78	58.66	3.10	5.45	58.34	0.52	0.91
Fast & Furious 7	24.27	27.66	3.39	14.00	25.78	1.51	6.22	25.57	1.3	5.36

5. Conclusion

Tackling the problems of ignoring consumer factors and low prediction accuracy in movie box office prediction, this paper proposes an adaptive attention movie box office prediction model with consumer sentinel. The experimental results show that the introduction of consumer data into the prediction model can improve the prediction accuracy on the basis of a movies own influencing factors. Compared with a single LSTM model and an LSTM model with an attention mechanism, the LSTM model with AAM has better prediction capabilities for movie box office prediction. In the future, the model can be further optimized by enriching the characteristics of expert experience, introducing more consumer characteristics, and adding movie reviews as an influencing factor.

Data Availability

The data used to support the findings of the study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," in *Proceedings of the International Conference on Computing, Communication & Automation*, pp. 933–937, Noida, India, May 2015.

[2] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, "Predicting movie Box-office revenues by exploiting large-scale social media content," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1509–1528, 2016.

[3] F. W. Derrick, N. A. Williams, and C. E. Scott, "A two-stage proxy variable approach to estimating movie box office receipts," *Journal of Cultural Economics*, vol. 38, no. 2, pp. 173–189, 2014.

[4] R. A. Nelson and R. Glotfelty, "Movie stars and box office revenues: an empirical analysis," *Journal of Cultural Economics*, vol. 36, no. 2, pp. 141–166, 2012.

[5] J. Prieto-Rodriguez, F. Gutierrez-Navratil, and V. Ateca-Amestoy, "Theatre allocation as a distributor's strategic variable over movie runs," *Journal of Cultural Economics*, vol. 39, no. 1, pp. 65–83, 2015.

[6] J. Du, H. Xu, and X. Huang, "Box office prediction based on microblog," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1680–1689, 2014.

[7] L. Zhang, J. Luo, and S. Yang, "Forecasting box office revenue of movies with BP neural network," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6580–6587, 2009.

[8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[9] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, Portland, OR, USA, September 2012.

[10] C. Yang, W. Jiang, and D. Shao, "Movie box office prediction algorithm based on LSTM model," *Data Communication*, vol. 5, p. 9, 2019.

[11] P. Marshall, M. Dockendorff, and S. Ibáñez, "A forecasting system for movie attendance," *Journal of Business Research*, vol. 66, no. 10, pp. 1800–1806, 2013.

[12] J. Hartzel, A. Agresti, and B. Caffo, "Multinomial logit random effects models," *Statistical Modelling: An International Journal*, vol. 1, no. 2, pp. 81–102, 2001.

[13] D. Barman, N. Chowdhury, and R. K. Singha, "To predict possible profit/loss of a movie to be launched using MLP with back-propagation learning," in *Proceedings of the 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)*, pp. 322–325, Kolkata, India, December 2012.

[14] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[15] Y. Zhou, L. Zhang, and Z. Yi, "Predicting movie box-office revenues using deep neural networks," *Neural Computing and Applications*, vol. 31, no. 6, pp. 1855–1865, 2019.

[16] H. Tahmasebi, R. Ravanmehr, and R. Mohamadrezaei, "Social movie recommender system based on deep autoencoder network using Twitter data," *Neural Computing and Applications*, pp. 1–17, 2020.

[17] S. Kim and M. Kang, "Financial series prediction using attention LSTM," 2019, http://arxiv.org/abs//1902.10877.

[18] J. Li, S. X. Pan, L. Huang, and X. Zhu, "A machine learning based method for customer behavior prediction," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1670–1676, 2019.
[19] L. Wang, Z. Hao, X. Han, and R. Zhou, "Gravity theory-based affinity propagation clustering Algorithm and its Applications," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 4, pp. 1125–1135, 2018.
[20] J. Niresh, N. Archana, and G. Anand Raj, "Optimisation of linear passive suspension system using MOPSO and design of predictive tool with artificial neural network," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 105–110, 2019.
[21] X. Fan, H. Lin, L. Yang et al., "Phonetics and ambiguity comprehension gated attention network for humor recognition," *Complexity*, vol. 2020, Article ID 2509018, 2020.
[22] N. Wang, L. Yang, Y. Zheng, X. Cai, X. Mei, and H. Dai, "A tri-attention neural network model-based recommendation," *Complexity*, vol. 2020, Article ID 3857871, 2020.

Hindawi
Complexity
Volume 2020, Article ID 6622927, 10 pages
https://doi.org/10.1155/2020/6622927

Research Article

A CNN-LSTM-Based Model to Forecast Stock Prices

Wenjie Lu,[1,2] Jiazheng Li,[3] Yifan Li,[3] Aijun Sun ,[1] and Jingyang Wang[3]

[1]*Business School, Jiangsu Second Normal University, Nanjing 210000, China*
[2]*School of Economics and Management, Hebei University of Science and Technology, Shijiazhuang 050018, China*
[3]*School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China*

Correspondence should be addressed to Aijun Sun; saj1970@jssnu.edu.cn

Received 22 October 2020; Revised 2 November 2020; Accepted 3 November 2020; Published 24 November 2020

Academic Editor: Abd E. I.-Baset Hassanien

Stock price data have the characteristics of time series. At the same time, based on machine learning long short-term memory (LSTM) which has the advantages of analyzing relationships among time series data through its memory function, we propose a forecasting method of stock price based on CNN-LSTM. In the meanwhile, we use MLP, CNN, RNN, LSTM, CNN-RNN, and other forecasting models to predict the stock price one by one. Moreover, the forecasting results of these models are analyzed and compared. The data utilized in this research concern the daily stock prices from July 1, 1991, to August 31, 2020, including 7127 trading days. In terms of historical data, we choose eight features, including opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and change. Firstly, we adopt CNN to efficiently extract features from the data, which are the items of the previous 10 days. And then, we adopt LSTM to predict the stock price with the extracted feature data. According to the experimental results, the CNN-LSTM can provide a reliable stock price forecasting with the highest prediction accuracy. This forecasting method not only provides a new research idea for stock price forecasting but also provides practical experience for scholars to study financial time series data.

1. Introduction

The change trend of the stock price has always been identified as a very important problem in the economic field [1]. Stock prices are affected by various internal and external factors, such as domestic and foreign economic environment, international situation, industry prospect, financial data of listed companies, and stock market operation. Thus, the forecasting method also has different emphasis [2, 3].

The traditional analysis method is based on economics and finance, which mainly uses the fundamental analysis method and the technical analysis method. On the one hand, the fundamental analysis method pays more attention to the intrinsic value of stocks and qualitatively analyzes the external factors that affect the stock, such as interest rate, exchange rate, inflation, industrial policy, finance of listed companies, international relations, and other economic and political factors. On the other hand, the technical analysis method mainly focuses on the direction of stock price, trading volume, and investors' psychological expectation, which primarily focuses on analyzing the stock index trajectory of individual stocks or the whole market by using *K*-line chart and other tools. At present, traditional fundamental analysis and technical analysis are still the most commonly employed methods for many organizations and individual investors [4, 5].

The accuracy of the traditional fundamental analysis method is difficult to be convincing. The reason is not only that the influencing factors are in a long-term cycle, but also the forecasting results are more dependent on the professional quality of analysts. As a financial time series, stock data have the characteristics of random walk [6]. Based on statistics and probability theory, some scholars use time series linear forecasting model to predict the short-term stock price with a large number of long-term data, such as vector autoregression (VAR) [7], Bayesian vector autoregression (BVAR) model [8], autoregressive integrated moving average mode (ARIMA) [9], and generalized autoregressive conditional heteroskedasticity model (GARCH) [10]. However, the accuracy of using time series

model alone is questioned due to the uncertainty and high noise characteristics of financial time series and the relationship between independent variables and dependent variables is prone to dynamic changes over time, which limits its further application and expansion [11].

It has certain limitations to predict stock price trend with single simply using the linear time series forecasting model or neural network model. At present, combining the advantages of various methods and using various best algorithms to improve the hybrid method is the development trend of financial time series deep learning [12]. Therefore, in order to make the best of the time series characteristics of data series, deeply mine the data features, and improve the accuracy of stock price forecasting, this paper proposes a stock price forecasting method based on CNN-LSTM for the stock closing price of the next day forecasting. Combining the advantages of convolutional neural networks (CNN) that can extract effective features from the data, and long short-term memory (LSTM) which can not only find the interdependence of data in time series data, but also automatically detect the best mode suitable for relevant data, this method can effectively improve the accuracy of stock price forecasting. The CNN-LSTM model uses CNN to extract the features of the input time data and uses LSTM to predict the stock closing price on the next day. In order to verify the effectiveness of the model, this paper uses the daily transaction data of 7127 trading days from July 1, 1991, to August 31, 2020, in which the first 6627 trading days data are the training set and the last 500 trading days data are the test set.

2. Related Work

At present, the financial market is a noisy, nonparametric dynamic system, and there are two main kinds of forecasting methods for stock price: traditional analysis method and machine learning method [13]. The traditional econometric methods or equations with parameters are not suitable for analyzing complex, high-dimensional, and noisy financial series data. In recent years, neural network has become a hot research direction in the field of stock forecasting because it can extract data features from a large number of high-frequency raw data without relying on prior knowledge. In 1988, White used neural network to predict IBM stock, but the experimental results were not good [14]. In 2003, Zhang used neural network and autoregressive integrated moving average model (ARIMA) to forecast stocks, respectively. The experimental results show that neural network has obvious advantages in nonlinear data forecasting, but the accuracy still needs to be improved [15]. In 2005, Sun et al. proposed a time series forecasting method based on neural network. This method combines the optimal partition algorithm (OPA) and radial basis function (RBF) neural network [16]. In 2014, Adhikari et al. proposed a method combining random walk (RW) and artificial neural network (ANN) to predict four financial time series data, and the results showed that the forecasting accuracy had a certain improvement [17]. In 2018, Zhang et al. proposed the network structure of stock price forecasting based on LM-BP neural network, which improved the traditional BP neural network training algorithm's shortcomings of slow training speed and low precision [18]. In 2018, the experimental results of Hu et al. show that convolutional neural network can predict time series, and deep learning is more suitable for solving the problem of time series. However, because CNN is more commonly used to solve image recognition and feature extraction, the forecasting accuracy of CNN alone is relatively low [19]. In 2020, Kamalov used MLP, CNN, and LSTM to forecast the stock price of four major US public companies. Experimental results showed that these three methods showed better results compared to similar studies that forecast the direction of price change [20]. In 2020, Xue et al. established a high-precision short-term forecasting model of financial market time series based on LSTM deep neural network and compared with the BP neural network, the traditional RNN, and the improved LSTM deep neural network. The results showed that the LSTM deep neural network has high forecasting accuracy and can effectively predict the time series of the stock market [21].

The main contributions of this paper are as follows:

(1) By analyzing the correlation and time series of stock price data, a new deep learning method (CNN-LSTM) is proposed to predict the stock price. In this method, CNN is used to extract the time feature of data, and LSTM is used for data forecasting. It can make full use of the time sequence of stock price data to obtain more reliable forecasting.

(2) By comparing the evaluation indexes of CNN-LSTM with multilayer perceptron (MLP), CNN, RNN, LSTM, and CNN-RNN, it is proved that CNN-LSTM has high forecasting accuracy and is more suitable for stock price forecasting.

3. CNN-LSTM

3.1. CNN-LSTM Model. CNN has the characteristic of paying attention to the most obvious features in the line of sight, so it is widely used in feature engineering. LSTM has the characteristic of expanding according to the sequence of time, and it is widely used in time series. According to the characteristics of CNN and LSTM, a stock forecasting model based on CNN-LSTM is established. The model structure diagram is shown in Figure 1, and the main structure is CNN and LSTM, including input layer, one-dimensional convolution layer, pooling layer, LSTM hidden layer, and full connection layer.

3.2. CNN. CNN is a network model proposed by Lecun et al. in 1998 [22]. CNN is a kind of feedforward neural network, which has good performance in image processing and natural language processing [23]. It can be effectively applied to the forecasting of time series. The local perception and weight sharing of CNN can greatly reduce the number of parameters, thus improving the efficiency of model learning [24]. CNN is mainly composed of two parts: convolution layer and pooling layer. Each convolution layer contains a plurality of convolution kernels, and its calculation formula is shown in formula (1). After the convolution operation of

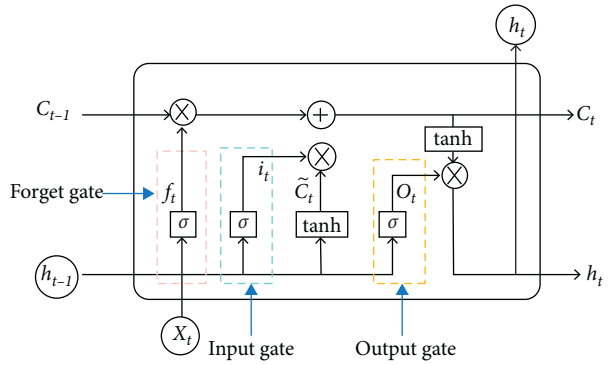


Figure 1: CNN-LSTM structure diagram.

the convolution layer, the features of the data are extracted, but the extracted feature dimensions are very high, so in order to solve this problem and reduce the cost of training the network, a pooling layer is added after the convolution layer to reduce the feature dimension:

$$l_t = \tanh\left(x_t * k_t + b_t\right), \tag{1}$$

where l_t represents the output value after convolution, tanh is the activation function, x_t is the input vector, k_t is the weight of the convolution kernel, and b_t is the bias of the convolution kernel.

3.3. LSTM. LSTM is a network model proposed by Schmidhuber et al. in 1997 [25]. LSTM is a network model designed to solve the longstanding problems of gradient explosion and gradient disappearance in RNN [26, 27]. It has been widely used in speech recognition, emotional analysis, and text analysis, as it has its own memory and can make relatively accurate forecasting [28, 29]. In recent years, it has also been adopted in the field of stock market forecasting [30–32]. There is only one repeating module in a standard RNN, and its internal structure is simple. It is usually a tanh layer. However, four of the LSTM modules are similar to the standard RNN modules, and they operate in a special interactive manner [33, 34]. The LSTM memory cell consists of three parts: the forget gate, the input gate, and the output gate, as shown in Figure 2.

The LSTM calculation process is as follows:

(1) The output value of the last moment and the input value of the current time are input into the forget gate, and the output value of the forget gate is obtained after calculation, as shown in the following formula:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \tag{2}$$

where the value range of f_t is (0,1), W_f is the weight of the forget gate, and b_f is the bias of the forget gate, x_t is the input value of the current time, and h_{t-1} is the output value of the last moment.

(2) The output value of the last time and the input value of the current time are inputted into the input gate, and the output value and candidate cell state of the input gate are obtained after calculation, as shown in the following formulas:

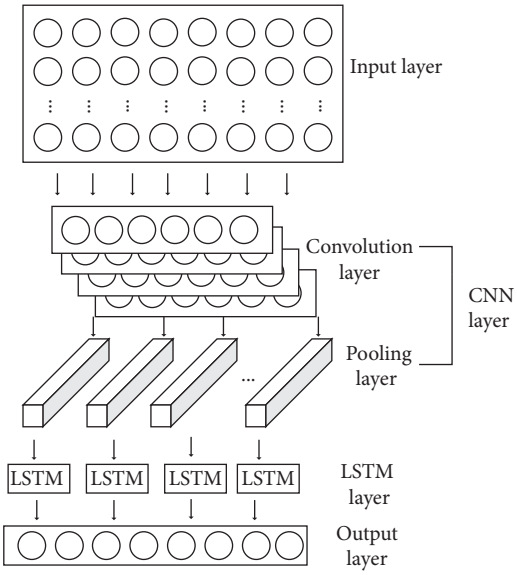


Figure 2: Architecture of LSTM memory cell.

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \tag{3}$$

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \tag{4}$$

where the value range of i_t is (0,1), W_i is the weight of the input gate, b_i is the bias of the input gate, W_c is the weight of the candidate input gate, and b_c is the bias of the candidate input gate.

(3) Update the current cell state as follows:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{5}$$

where the value range of C_t is (0,1).

(4) The output h_{t-1} and input x_t are received as input values of the output gate at time t, and the output o_t of the output gate is obtained as follows:

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right), \tag{6}$$

where the value range of o_t is (0,1), W_o is the weight of the output gate, and b_o is the bias of the output gate.

(5) The output value of LSTM is obtained by calculating the output of the output gate and the state of the cell, as shown in the following formula

$$h_t = o_t * \tanh\left(C_t\right). \tag{7}$$

3.4. CNN-LSTM Training and Prediction Process. The CNN-LSTM process of training and prediction is shown in Figure 3.

The main steps are as follows:

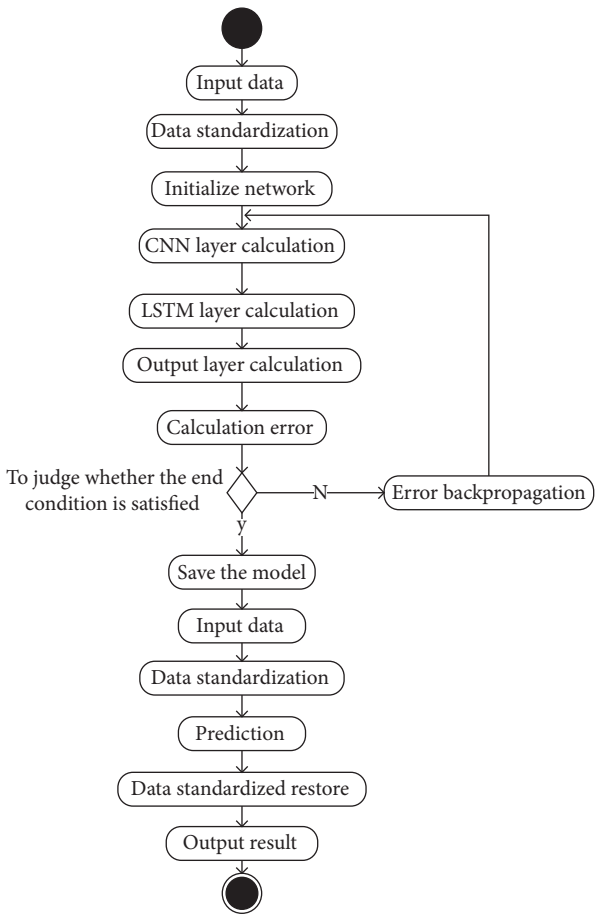


Figure 3: Activity diagram of CNN-LSTM training and prediction process.

(1) Input data: input the data required for CNN-LSTM training.

(2) Data standardization: as there is a large gap in the input data, in order to train the model better, the z-score standardization method is adopted to standardize the input data, as shown in the following formula:

$$y_i = \frac{x_i - \overline{x}}{s}, \tag{8}$$

$$x_i = y_i * s + \overline{x}, \tag{9}$$

where y_i is the standardized value, x_i is the input data, $\overline{x}$ is the average of the input data, and s is the standard deviation of the input data.

(3) Initialize network: initialize the weights and biases of each layer of the CNN-LSTM.

(4) CNN layer calculation: the input data are successively passed through the convolution layer and pooling layer in the CNN layer, the feature extraction of the input data is carried out, and the output value is obtained.

(5) LSTM layer calculation: the output data of the CNN layer are calculated through the LSTM layer, and the output value is obtained.

(6) Output layer calculation: the output value of the LSTM layer is input into the full connection layer to get the output value.

(7) Calculation error: the output value calculated by the output layer is compared with the real value of this group of data, and the corresponding error is obtained.

(8) To judge whether the end condition is satisfied: the conditions for the end are to complete a predetermined number of cycles, the weight is lower than a certain threshold, and the error rate of the forecasting is lower than a certain threshold. If one of the conditions for the end is met, the training will be completed, update the entire CNN-LSTM network, and go to step 10; otherwise, go to step 9.

(9) Error backpropagation: propagate the calculated error in the opposite direction, update the weight and bias of each layer, and go to step 4 to continue to train the network.

(10) Save the model: save the trained model for forecasting.

(11) Input data: input the input data required for the forecasting.

(12) Data standardization: the input data are standardized according to formula (8).

(13) Forecasting: input the standardized data into the trained model of CNN-LSTM, and then get the corresponding output value.

(14) Data standardized restore: the output value obtained through the model of CNN-LSTM is the standardized value, and the standardized value is restored to the original value. As shown in the following formula (9).

where x_i is the standardized restored value, y_i is the output value of the CNN-LSTM, s is the standard deviation of the input data, and $\overline{x}$ is the average value of the input data.

(15) Output result: output the restored results to complete the forecasting process.

4. Experiments

In order to prove the effectiveness of CNN-LSTM, we compared this method with MLP, CNN, RNN, LSTM, and CNN-RNN using the same training set and test set data under the same operating environment. All the experiments are carried out under the running environment of Intel i7-4700H 2.6 GHz, 12 GBs of RAM, 500 GBs of hard disk and Windows 10. According to the influence factors, including the opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and change, the next day's closing price is predicted.

Table 1: Partial sample data.

Date	Opening price	Highest price	Lowest price	Closing price	Volume (share)	Turnover (RMB)	Ups and downs	Change (%)
1991/7/1	136.64	138.62	136.56	136.85	2294000	12469884	−0.71	−0.5161
1991/7/2	135.91	135.96	135.69	135.96	283800	3794100	−0.89	−0.6503
1991/7/3	135.28	135.96	134.98	135.27	271500	1818504	−0.69	−0.5075
1991/7/4	136.63	136.63	134.19	136.63	1339400	8095138	1.36	1.0054
1991/7/5	136.01	137.68	135.9	135.96	1454000	9394861	−0.67	−0.4904

4.1. Data. In this experiment, the Shanghai Composite Index (000001) is selected as the experimental data. The daily trading data of 7127 trading days from July 1, 1991, to August 31, 2020, are obtained from the wind database. Each piece of data contains eight items, namely, opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and change. Some of the data are shown in Table 1. Take the data of the first 6627 trading days as training set and the data of the last 500 trading days as test set.

4.2. Model Implementation. In order to evaluate the forecasting effect of CNN-LSTM, the mean absolute error (MAE), root mean square error (RMSE), and R-square (R^2) are used as the evaluation criteria of the methods.

The MAE calculation formula is as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y}_i - ty_i\right|, \tag{10}$$

where $\widehat{y}_i$ is the predictive value and y_i is the true value. The smaller the value of MAE, the better the forecasting.

The RMSE calculation formula is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}, \tag{11}$$

where $\widehat{y}_i$ is the predictive value and y_i is the true value. The smaller the value of RMSE, the better the forecasting.

The R^2 calculation formula is as follows:

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2\right)/n}{\left(\sum_{i=1}^{n}\left(\bar{y}_i - t\widehat{y}_i\right)^2\right)/n,} \tag{12}$$

where $\widehat{y}_i$ is the predictive value, y_i is the true value, and $\overline{y_i}$ is the average value. The value range of R^2 is (0,1).

The closer the value of MAE and RMSE to 0, the smaller the error between the predicted value and the real value, the higher the forecasting accuracy. The closer R^2 is to 1, the better the fitting degree of the model is.

4.3. Implementation of CNN-LSTM. The parameter setting of the CNN-LSTM for this experiment is shown in Table 2.

According to the parameter setting of CNN-LSTM network, we can know that the specific model is constructed as follows: the input training set data is a three-dimensional data vector *(None, 10, 8)*, in which 10 is the size of the time_step and 8 is the 8 features of the input dimension. First, the data enter the one-dimensional convolution layer to further extract features and obtain a three-dimensional output vector *(None, 10, 32)*, in which 32 is the size of the convolution layer filters. Next, the vector enters the pooling layer, and a three-dimensional output vector *(None, 10, 32)* is also obtained. And then, the output vector enters the LSTM layer for training, and the output data *(None, 64)* after training enter another layer of full connection layer to get the

Table 2: Parameter setting of CNN-LSTM.

Parameters	Value
Convolution layer filters	32
Convolution layer kernel_size	1
Convolution layer activation function	tanh
Convolution layer padding	Same
Pooling layer pool_size	1
Pooling layer padding	Same
Pooling layer activation function	Relu
Number of hidden units in LSTM layer	64
LSTM layer activation function	tanh
Time_step	10
Batch_size	64
Learning rate	0.001
Optimizer	Adam
Loss function	mean_absolute_error
Epochs	100

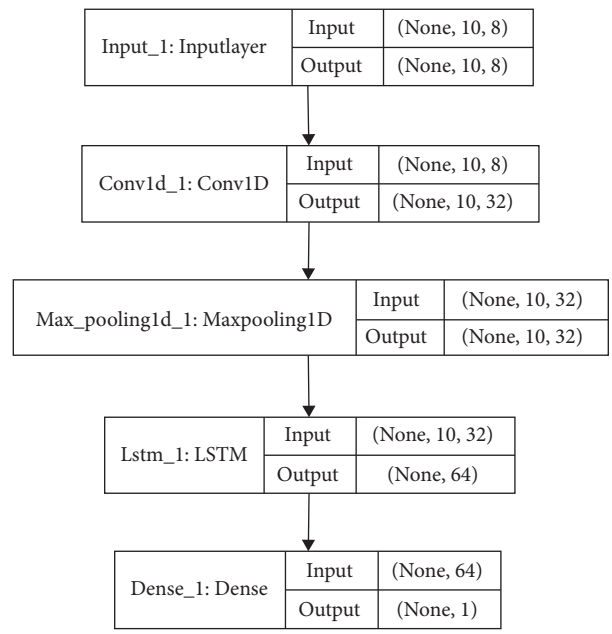


Figure 4: The model structure of CNN-LSTM.

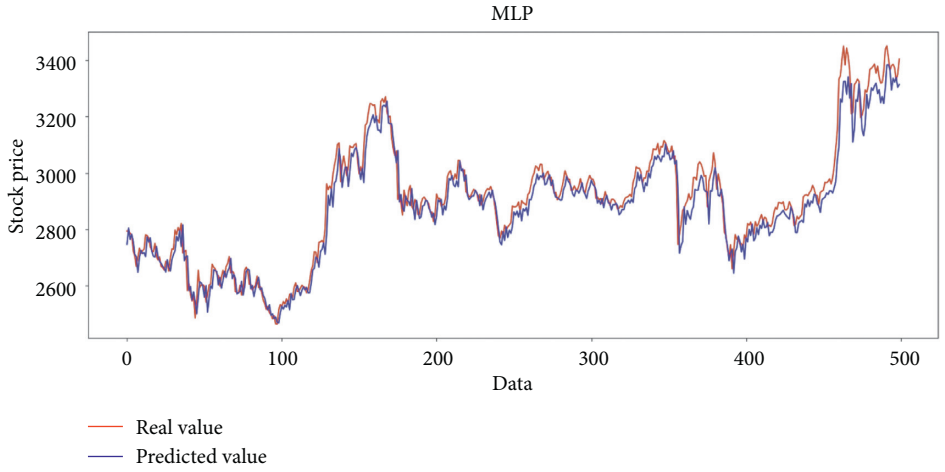


Figure 5: Comparison of the predicted value and the real value for MLP.

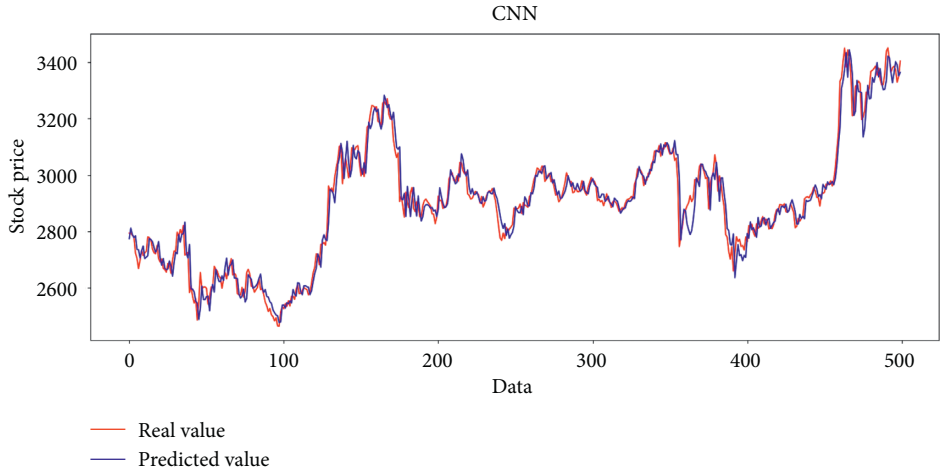


Figure 6: Comparison of the predicted value and the real value for CNN.

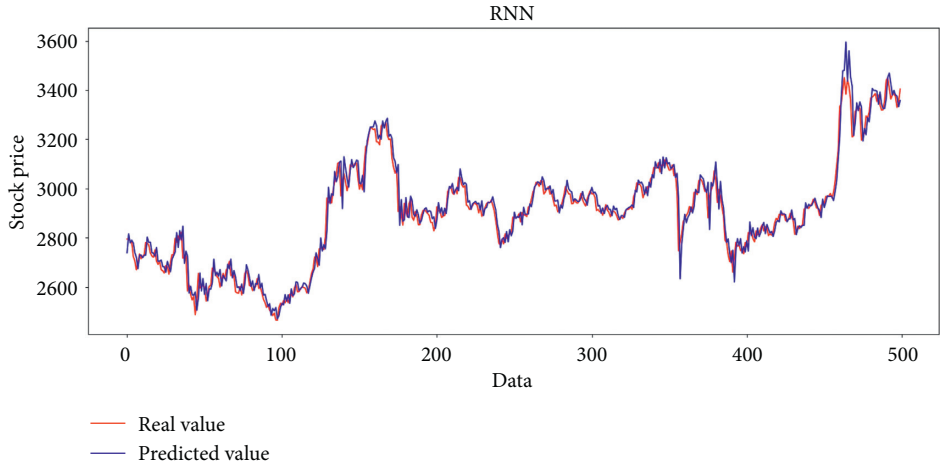


Figure 7: Comparison of the predicted value and the real value for RNN.

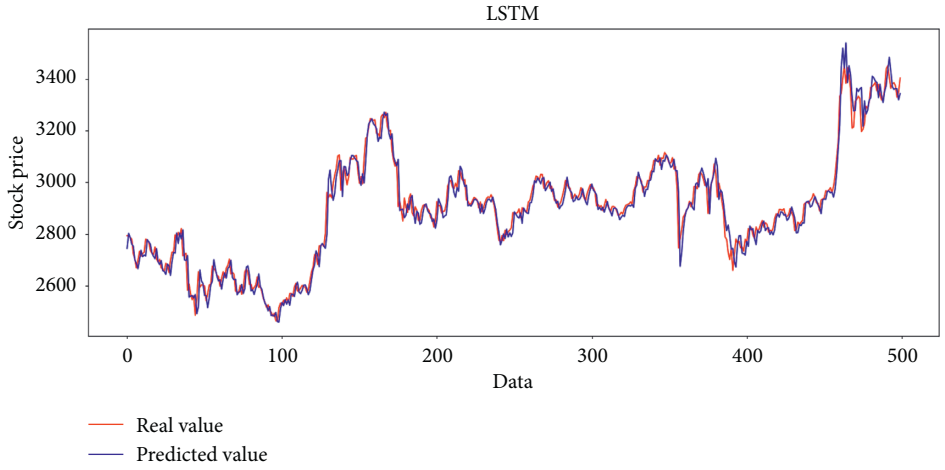


Figure 8: Comparison of the predicted value and the real value for LSTM.

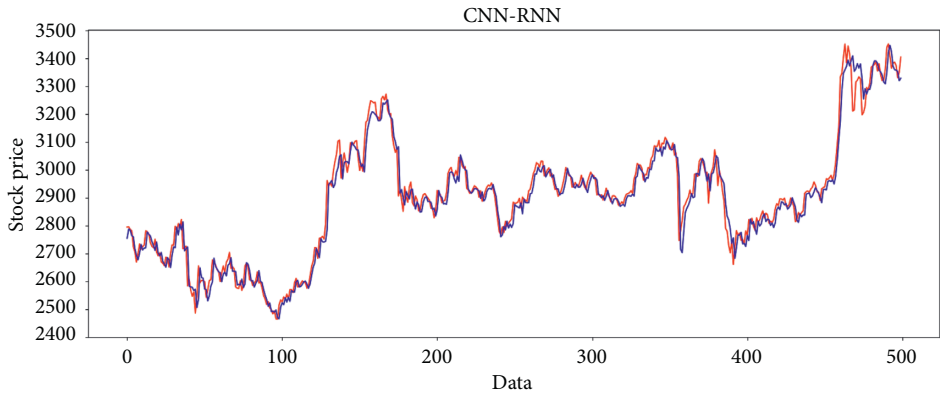


Figure 9: Comparison of the predicted value and the real value for CNN-RNN.

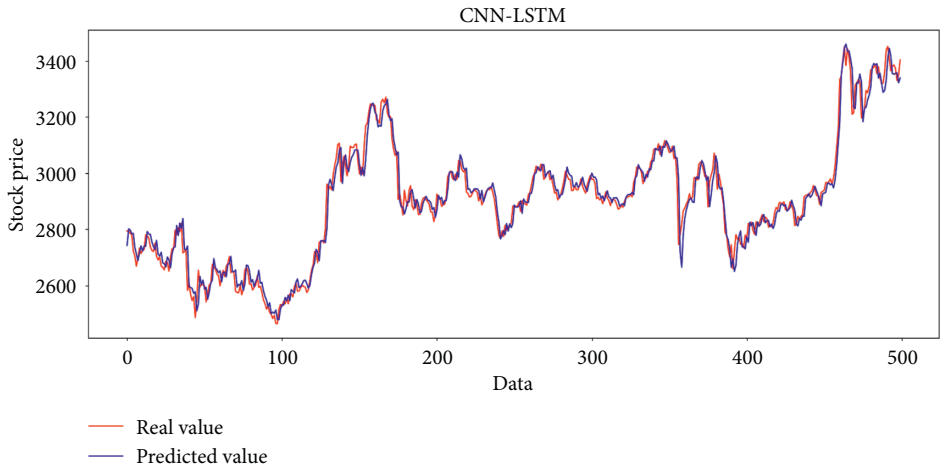


Figure 10: Comparison of the predicted value and the real value for CNN-LSTM.

Table 3: Comparison of nine methods evaluation indexes.

Method	MAE	RMSE	R^2
MLP	37.584	49.799	0.9442
CNN	30.138	42.967	0.9585
RNN	29.916	42.957	0.9593
LSTM	28.712	41.003	0.9622
CNN-RNN	28.285	40.538	0.9630
CNN-LSTM	**27.564**	**39.688**	**0.9646**

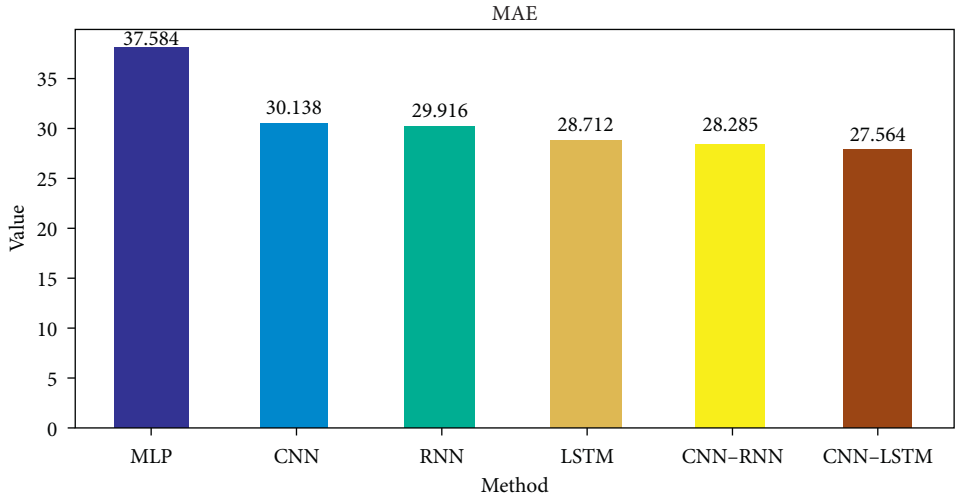


Figure 11: The result of MAE comparison among different methods.

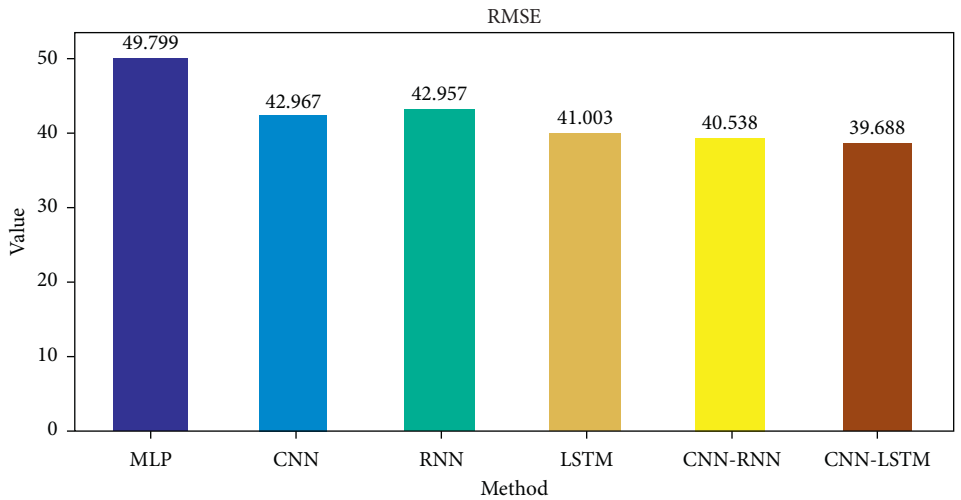


Figure 12: The result of RMSE comparison among different methods.

output value; 64 is the number of hidden units in the LSTM layer. The specific CNN-LSTM model structure is shown in Figure 4.

5. Results

After using the processed training set data to train MLP, CNN, RNN, LSTM CNN-RNN, and CNN-LSTM, respectively, the model completed by training is used to predict the test set data, and the real value is compared with the predicted value as shown in Figures 5–10.

In Figures 5–10, among the six forecasting methods, the broken line fitting degree of real value and predicted value is CNN-LSTM, CNN-RNN, LSTM, CNN, RNN, and MLP. CNN-LSTM has the highest degree of broken line fitting which almost coincides with each other, and MLP has the lowest degree of broken line fitting.

According to the predicted value and real value of each method, the evaluation index of each method can be

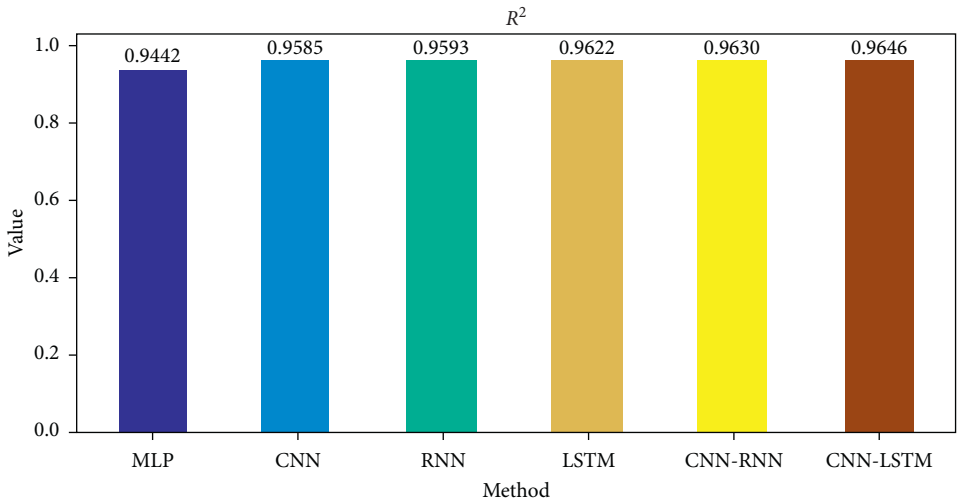


Figure 13: The result of R^2 comparison among different methods.

calculated, and the comparison results of the six methods are shown in Table 3 and Figures 11–13.

From Table 3 and Figures 10–12, the MAE and RMSE of MLP are the largest and R^2 is the smallest, while the MAE and RMSE of CNN-LSTM are the smallest, R^2 is the largest, and the closest is 1.

By comparing LSTM with RNN, the MAE and RMSE of LSTM decrease, R2 increases by 0.3%, MAE decreases from 29.916 to 28.712 by 4.0%, and RMSE decreases from 42.957 to 41.003 by 4.5%, so LSTM was better than RNN. However, the error measurement indexes MAE and RMSE of CNN-LSTM are the smallest, and the maximum R^2 is close to 1. Compared with LSTM, after CNN layer, MAE and RMSE of CNN-LSTM proposed in this paper are lower than those without CNN layer; R^2 has a certain improvement; MAE decreases by 4.0%, from 28.712 to 27.564; RMSE decreases by 3.2%, from 41.003 to 39.688; and R^2 increases by 0.2%. It shows that the forecasting performance of LSTM can be effectively improved by extracting data features through CNN.

The results show that the performance of CNN-LSTM is the best among the six methods. In terms of forecasting accuracy, MAE is 27.564 and RMSE is 39.688, which is the smallest among the six forecasting models and has high forecasting accuracy, in terms of forecasting performance, and the R^2 of CNN-LSTM is 0.9646, which is improved by 2.2%, 0.6%, 0.5%, and 0.2%, respectively, compared with the other four methods. Therefore, the CNN-LSTM proposed in this paper is superior to the other four comparative models in terms of fitting degree and error value. It can well predict the closing price of the next day and provide a reference for investors' investment.

6. Conclusions

According to the chronological characteristics of stock price data, this paper proposes a CNN-LSTM to predict the stock closing price of the next day. The method uses opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and change of the stock data as the input, making full use of the time sequence characteristics of the stock data. CNN is used to extract the features of the input data. LSTM is used to learn the extracted feature data and predict the closing price of the stock the next day. This paper takes the relevant data of the Shanghai Composite Index as an example to verify the experimental results. The experimental results show that the CNN-LSTM has the highest forecasting accuracy and the best performance compared with the MLP, CNN, RNN, LSTM, and CNN-RNN. MAE and RMSE are the smallest of all methods, and R^2 is close to 1. CNN-LSTM is suitable for the forecasting of stock prices and can provide a relevant reference for investors to maximize investment returns. CNN-LSTM also provides the proposal of practical experience for people's research on financial time series data. However, the model still has some shortcomings. For example, it only considers the impact of stock price data on closing prices and fails to integrate emotional factors such as news and national policy into the forecast. Our future research work is mainly to increase the sentiment analysis of stock-related news and national policies, so as to ensure the accuracy of stock forecast.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was funded by the Soft Science Project of Hebei Province, Grant 205576142D, and Humanities and Social Science Research Project of Hebei Education Department, Grant SD201010.

References

[1] R. Vanaga and B. Sloka, "Financial and capital market commission financing: aspects and challenges," *Journal of Logistics, Informatics and Service Science*, vol. 7, no. 1, pp. 17–30, 2020.

[2] L. Zhang and H. Kim, "The influence of financial service characteristics on use intention through customer satisfaction with mobile fintech," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 82–94, 2020.

[3] L. Badea, V. Ionescu, and A.-A. Guzun, "What is the causal relationship between stoxx europe 600 sectors? But between large firms and small firms?" *Economic Computation And Economic Cybernetics Studies And Research*, vol. 53, no. 3, pp. 5–20, 2019.

[4] J. Sousa, J. Montevechi, and R. Miranda, "Economic lot-size using machine learning, parallelism, metaheuristic and simulation," *Journal of Logistics, Informatics and Service Science*, vol. 18, no. 2, pp. 205–216, 2019.

[5] A. Coser, M. M. Maer-Matei, and C. Albu, "Predictive models for loan default risk assessment," *Economic Computation And Economic Cybernetics Studies And Research*, vol. 53, no. 2, pp. 149–165, 2019.

[6] R. Qiao, "Stock prediction model based on neural network," *Operations Research and Management Science*, vol. 28, no. 10, pp. 132–140, 2019.

[7] C. Jung and R. Boyd, "Forecasting UK stock prices," *Applied Financial Economics*, vol. 6, no. 3, pp. 279–286, 1996.

[8] W. Bleesser and P. Liicoff, "Predicting stock returns with bayesian vector autoregressive," *Data Analysis, Machine Learning and Applications*, vol. 1, pp. 499–506, 2005.

[9] A. Adebiyi, A. Adewumi, and C. Ayo, "Stock price prediction using the ARIMA model," in *Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, IEEE, Cambridge, UK, March 2014.

[10] C. Zhang, X. Cheng, and M. Wang, "An empirical research in the stock market of Shanghai by GARCH model," *Operations Research and Management Science*, vol. 4, pp. 144–146, 2005.

[11] Q. Yang and C. Wang, "A study on forecast of global stock indices based on deep LSTM neural network," *Statistical Research*, vol. 36, no. 6, pp. 65–77, 2019.

[12] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation And Economic Cybernetics Studies And Research*, vol. 53, no. 2, pp. 77–92, 2019.

[13] J. Li, S. Pan, L. Huang, and X. Zhu, "A machine learning based method for customer behavior prediction," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1670–1676, 2019.

[14] H. White, "Economic prediction using neural networks: the case of IBM daily stock returns," *Earth Surface Processes & Landforms*, vol. 8, no. 5, pp. 409–422, 1988.

[15] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, no. 1, pp. 159–175, 2003.

[16] Y. Sun, Y. Liang, and W. Zhang, "Optimal partition algorithm of the RBF neural network and its application to financial time series forecasting," *Neural Computing and Applications*, vol. 14, pp. 1441–1449, 2005.

[17] R. Adhikari and R. Agrawal, "A combination of artificial neural network and random walk models for financial time series forecasting," *Neural Computing and Applications*, vol. 24, pp. 305–315, 2014.

[18] L. Zhang, F. Wang, B. Xu, W. Chi, Q. Wang, and T. Sun, "Prediction of stock prices based on LM-BP neural network and the estimation of overfitting point by RDCI," *Neural Computing and Applications*, vol. 30, no. 5, pp. 1425–1444, 2018.

[19] Y. Hu, "Stock market timing model based on convolutional neural network – a case study of Shanghai composite index," *Finance& Economy*, vol. 4, pp. 71–74, 2018.

[20] E. Alibasic, B. Fazo, and I. Petrovic, "A new approach to calculating electrical energy losses on power lines with a new improved three-mode method," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 2, pp. 405–411, 2019.

[21] Y. Xue, C. Wang, and C. Miao, "Research on financial assets transaction prediction model based on LSTM neural network," *Neural Computing and Applications*, vol. 1, 2020.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] B. S. Kim and T. G. Kim, "Cooperation of simulation and data model for performance analysis of complex systems," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 608–619, 2019.

[24] L. Qin, N. Yu, and D. Zhao, "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 2, pp. 528–535, 2018.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *MIT Press*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] V.-D. Ta, C.-M. Liu, and D. A. Tadesse, "Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading," *Applied Sciences*, vol. 10, no. 2, pp. 437–457, 2020.

[27] O. Zarrad, M. A. Hajjaji, and M. N. Mansouri, "Hardware implementation of hybrid wind-solar energy system for pumping water based on artificial neural network controller," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 35–44, 2019.

[28] T. Saric, G. Simunovic, D. Vukelic, K. Simunovic, and R. Lujic, "Estimation of CNC grinding process parameters using different neural networks," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 6, pp. 1770–1775, 2018.

[29] N. Gupta and A. Jalal, "Integration of textual cues for fine-grained image captioning using deep CNN and LSTM," *Neural Computing and Applications*, vol. 12, pp. 1–10, 2019.

[30] A. Yadav, C. K. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," *Procedia Computer Science*, vol. 167, pp. 2091–2100, 2020.

[31] H. Y. Kim and C. H. Won, "Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models," *Expert Systems with Applications*, vol. 103, pp. 25–37, 2018.

[32] N. C. Petersen, R. Christoffer, F. Rodrigues, and F. C. Pereira, "Multi-output bus travel time prediction with convolutional LSTM neural network," *Expert Systems with Applications*, vol. 120, pp. 426–435, 2019.

[33] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications*, vol. 32, no. 13, pp. 9713–9729, 2020.

[34] B. Svetlana and T. Ioannis, "An ensemble of LSTM neural networks for high-frequency stock market classification," *Journal of Forecasting*, vol. 38, no. 6, pp. 600–619, 2019.

Hindawi
Complexity
Volume 2020, Article ID 6650343, 9 pages
https://doi.org/10.1155/2020/6650343

Research Article

Multiple Channel Integration Quality Assessment Method Using NARX

Xiaolei Wang [1] **and Yingzhao He** [2]

[1]*School of Economic and Management, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Business, Northwest Normal University, Lanzhou 730070, China*

Correspondence should be addressed to Xiaolei Wang; 12113148@bjtu.edu.cn

Received 20 October 2020; Revised 1 November 2020; Accepted 7 November 2020; Published 21 November 2020

Academic Editor: Abd E. I.-Baset Hassanien

To improve the accuracy of the multiple channel integration quality (MCIQ) evaluation, this paper proposes a comprehensive evaluation method using the nonlinear autoregressive exogenous model (NARX) and constructs an index system. First, the entropy method is used to determine the objective weight of each indicator. The indicators used in this paper are process consistency, information consistency, emotional value, procedural value, service structure transparency, online result value, business relevance, and online purchase intention. Second, an improved gray relational analysis (GRA) algorithm is used to obtain the comprehensive gray relational degree between the above eight indicators' standard samples and the tested samples. Then, this study uses the dataset preprocessed with the GRA algorithm for training the NARX model. Then, this study uses the trained model to evaluate the quality of multiple channel integration comprehensively. Next, this study uses standardized methods to quantify the evaluation results to provide new ideas and theoretical guidance for teaching traditional retailers to use the advantages of multiple channels to expand their online business. This paper uses 50,000 consecutive samples of a product for 3 months as a dataset in the experimental part. Through the GRA method and the NARX model, the comprehensive gray relational degree between the test sample and the ideal sample is obtained, and the results are quantified. Experiments show that, compared with the GRA method, this paper's method has a higher degree of fit between the output value and the target value.

1. Introduction

With the rise and development of Internet companies, more and more "online shops" have emerged. Compared with traditional offline sales, online sales have gained more momentum in recent years, and there are even online "fitting rooms." To comprehensively analyze the influencing factors of consumers' purchase intentions and put forward valuable suggestions for traditional offline retail stores, we used the NARX neural network to conduct multichannel integration of customer purchasers' intentions and obtained corresponding results.

At present, many scholars have done some research on the quality of multiple channel integration. Berman et al. [1] believed that an excellent multiple channel integration strategy has the following characteristics: highly integrated promotion, cross-channel product consistency, shared customer integrated information system, cross-channel pricing and inventory information, online purchase and offline delivery service process, and provides multiple channel opportunity searches for suitable people. Saeed et al. [2] believed that the information system integration of retailers' virtual and physical channels can be divided into content integration, information integration, and logistics integration according to the value-added services provided. Chan et al. [3] showed that the systemic relationship between online and offline channels includes four dimensions: reinforcement, collaboration, interaction, and how much they complement each other. Lee et al. [4] divided the practice and effectiveness of retailers' cross-channel integration into five dimensions: information consistency, channel selection freedom, e-mail marketing effectiveness,

channel interaction, and store customer service evaluation. Oh et al. [5] divided the retail process integration into six dimensions: commodity and price information integration, promotion information integration, transaction information integration, information acquisition integration, order fulfillment integration, and customer service integration. There are two manifestations of retailers' multiple channel integration. From the process's perspective, multiple channel integration is the comprehensive cross-channel allocation and function optimization of various retail portfolio factors. The service structure level and interaction consistency of the channel have been improved. Sousa et al. [6] proposed the concept of MCIQ and believed that multiple channel service quality includes three components: physical, virtual, and integrated service quality (i.e., MCIQ). Among them, MCIQ reflects the overall result of the multiple channel integration framework. Through in-depth interviews with consumers, it is found that consumers are more likely to make explicit judgments about multiple channel integration [7]. Therefore, this study measures multiple channel integration from the perspective of results.

Recently, there are many classifications of quality evaluation methods for multiple channel integration, but they all have limitations, such as low accuracy and cumbersome calculation processes. To improve MCIQ evaluation's accuracy and overcome the main shortcomings of existing methods, this paper proposes a comprehensive evaluation method of MCIQ using the NARX model. This paper is mainly divided into four parts. Firstly, the objective weight value of each indicator is determined according to the entropy method. Secondly, the improved GRA method is used to obtain the comprehensive gray correlation degree. Third, the NARX model is trained using the dataset processed by the GRA method. Then, the trained NARX model is used to predict the comprehensive gray correlation degree. Finally, the results calculated by the GRA method and NARX model are standardized and compared with the target value, and the quantitative results of MCIQ are given.

2. Literature Review

2.1. Multiple Channel Integration Quality and Its Assessment Method. Sousa et al. [6] proposed the concept of MCIQ and believed that multiple channel service quality includes three components: physical, virtual, and integrated service quality. Among them, MCIQ reflects the overall result of the multiple channel integration framework. MCIQ is defined as the ability to provide customers with a seamless service experience through multiple channels. It includes two components: service structure and interactive integration. The service structure is the quality of customer combined services (or service elements) and related channels. Interaction integration refers to the consistency of interaction between customers and channels among different channels, leading to a consistent service experience.

There are many classifications of comprehensive evaluation methods. The relationship between evaluation and use of information characteristics can be divided into the data-driven evaluation, model-driven evaluation, expert knowledge-driven evaluation, and evaluations based on data, models, and expert knowledge [8]. According to the theoretical basis on which each evaluation method is based, comprehensive evaluation methods are roughly divided into four categories.

2.1.1. Experts Grading Method. The main methods are the expert meeting method, Delphi method, and Gl method [9]. Such methods are simple to operate and can use experts' knowledge, and the conclusions are easy to use. The disadvantage is that the subjectivity is relatively strong, and the conclusion is difficult to converge in the evaluation of multiple people. It is suitable for decision-making analysis objects at the strategic level, and systems that cannot be quantified are difficult to quantify [10].

2.1.2. Operations Research and Other Mathematical Methods

(1) Data Envelopment Analysis Method. Evaluating the relative effectiveness of units of the same type is based on sex versus efficiency, based on multitarget input and multitarget output, and based on a set of criteria to determine the frontier of significant production [11]. It is possible to evaluate large systems with multiple inputs and multiple outputs and use "window" technology to find unit weaknesses and improve. The disadvantage is that it only shows the evaluation unit's relative development index and cannot show the actual development level. Applicable evaluation objects include evaluating the technology, scale effectiveness of the production function in economics, industry benefit evaluation, and the effectiveness of the education sector.

(2) Analytic Hierarchy Process Method. The working principle of the method is to target a multilevel structure system, determine multiple judgment matrices by comparing relative quantities, take the characteristic phasor corresponding to their characteristic roots as weights, and finally synthesize the total weights and sort them [12]. It is a method with relatively high reliability, a small error, and a wide range of applications. It has promising service quality evaluation applications, cost-benefit decision-making, resource allocation sequence, and conflict analysis [13].

(3) Fuzzy Comprehensive Evaluation Method. The fuzzy comprehensive evaluation method's working principle is to replace the number in the judgment matrix of the analytic hierarchy process with a fuzzy set by introducing a membership function and quantify the constraint conditions to perform mathematical solutions [14]. This method can overcome the shortcomings of the "unique solution" in the traditional method and obtain multiple problem solutions based on different possibilities. It has scalability and conforms to the "flexible management" idea in modern management [15]. It is suitable for consumer preference identification, decision-making expert systems, securities investment analysis, bank project loan object identification, quality evaluation, and safety evaluation [16].

(4) Gray Comprehensive Evaluation Method. This method selects the optimal value from the evaluated object's various indicators as the evaluation standard. It then compares and ranks the evaluated objects using the degree of similarity between each plan and the optimal plan [17]. In the gray comprehensive evaluation method, the data do not need to be normalized, and there is no need for many samples. The calculation is simple, and the reliability is substantial [18]. The disadvantage is that the sample data must have time series characteristics and have all the disadvantages of "relative evaluation." It is suitable for projects that deal with inadequate information systems and only have a small amount of observation data.

2.1.3. Intelligent Evaluation Method. The intelligent evaluation method is represented by the artificial neural network evaluation method, an artificial neural network technology that simulates the human brain's intelligent processing process. It can "learn or be trained to acquire knowledge" through the BP algorithm. It stores the acquired knowledge in the neuron's weight and reproduces the relevant information through association [19]. It can "compute" and "refine" the objective laws of the evaluation object itself and to evaluate the evaluation objects of the same attribute [20]. The method's advantage is that the network has self-adaptability and fault tolerance and can handle large and complex systems with nonlinearity, nonlocality, and non-convexity. The disadvantage is that the accuracy is not high, and many training samples are required. Its application areas continue to expand, involving bank loan projects, stock price evaluation, and evaluation of the comprehensive urban development [21].

2.1.4. Statistical Methods

(1) Principal Component Analysis Method. This method's working principle is based on the common elements that dominate the existence of related economic variables. The goal is to study the correlation matrix of the original variables' internal structure and find several irrelevant comprehensive evaluation indicators that affect a specific economic process to represent the original variables. The principal component analysis method is comprehensive, comparable, objective, and reasonable. The disadvantage is that the actual meaning of new variables is often challenging to figure out, and the calculation process is cumbersome. It is suitable for evaluation items that classify evaluation objects, which have many evaluation indicators and have complex relationships between indicators.

(2) Cluster Analysis Method. Cluster analysis is an evaluation method for systematic clustering that calculates the distance between indicators or the similarity coefficient. It can solve the evaluation objects with a large degree of correlation. The disadvantage is that it requires a large amount of statistical data and does not reflect the objective development level. It is suitable for portfolio investment selection and regional development level evaluation.

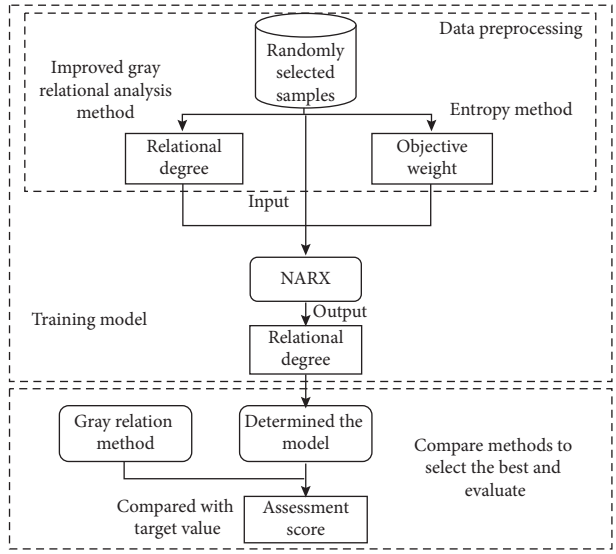


Figure 1: Methodology.

Although there are many mature methods and models for integrated quality evaluation, as mentioned above, they all have their limitations. This paper uses a NARX neural network to overcome low accuracy in general intelligent evaluation methods and the need for a large number of training samples. The improved GRA method and entropy method are used to preprocess the sample data so that the MCIQ can have higher accuracy.

3. Methodology

As shown in Figure 1, this research is mainly divided into four parts:

(1) Determine the objective weight value of each indicator based on the entropy method.

(2) The improved GRA method is used to obtain a comprehensive gray relational degree.

(3) The NARX model is trained with the dataset processed by the GRA method. Then, the trained NARX model is used to predict the comprehensive gray relational degree.

(4) The result calculated by the GRA method and NARX model is standardized and compared with the target value. Finally, the quantitative results of MCIQ are given.

Each indicator's objective weight can be obtained through the entropy method, which is based on the score and standardized as the target value. The improved GRA method is used to associate the standard sample of each evaluation indicator constructed with the sample to be tested. Then, the comprehensive gray relational degree of the index is obtained, and the improved comprehensive gray relational degree is used as the output of the NARX model.

Next, the evaluation index data in the sample to be tested are used as the input to the NARX model to train the NARX neural network. Then, the trained NARX model is used to predict and obtain the gray relational degree. Because each index's weight will be obtained in calculating the comprehensive gray relational degree, it can be standardized to obtain the sample quality evaluation score. Then, the results obtained by NARX and GRA are compared with the target value. The result closer to the target value is better and more accurate. Finally, the results of the MCIQ evaluation are obtained. These results provide new ideas and theoretical guidance for traditional retailers to use the advantages of multiple channels to expand their businesses.

4. Multiple Channel Integration Quality Evaluation Model Using NARX

4.1. Comprehensive Evaluation Index System Construction. This research divides the retailer's MCIQ into 8 dimensions from the perspective of customer perception, as shown in Table 1.

A comprehensive evaluation index for MCIQ contains the following dimensions: process consistency x_1, information consistency x_2, emotional value x_3, procedural value x_4, service structure transparency x_5, online result value x_6, business relevance x_7, and online purchase intention x_8.

4.2. MCIQ Data Processing

4.2.1. Improved GRA. The GRA is carried out by identifying the degree of similarity or difference between system factors' development trends. The GRA method can analyze the correlation between the standard samples of the above eight indicators and the samples to be evaluated to obtain the comprehensive gray relational degree of the indicators [22]. The main steps are as follows:

(1) Constructie the original evaluation matrix for gray correlation evaluation:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \\ \vdots \\ X_m \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{jn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn} \end{bmatrix}, \quad 1 \le i \le m,\ 1 \le j \le m, \tag{1}$$

where x_{ij} represents the j-th index of the i-th object and the original evaluation matrix is composed of m objects and n indexes. The reference number is $X_0 = \begin{bmatrix} X_{00} & X_{01} & \cdots & X_{0j} & \cdots & X_{0n} \end{bmatrix}$.

(2) Standardize data processing and establish an evaluation absolute difference matrix:

Calculate the absolute value of the difference between each element of the original evaluation matrix and its corresponding element in the reference sequence denoted as the following equation:

$$[\Delta_k] = (X_k - X_0), \quad k = 1, 2, \ldots, n. \tag{2}$$

(3) Calculate each index's gray relational degree between the sample to be evaluated and the standard sample [23] and transform the absolute difference matrix in equation (2). The correlation coefficient is shown in equation (2):

$$\xi_{ij} = \frac{\min \Delta_{i,j}\ {}_{1 \le j \le n}^{1 \le i \le m} + \rho \max \Delta_{i,j}\ {}_{1 \le j \le n}^{1 \le i \le m}}{\Delta_{i,j} + \rho \max \Delta_{i,j}\ {}_{1 \le j \le n}^{1 \le i \le m}}. \tag{3}$$

In equation (3), ρ is the resolution coefficient.

This paper adopts the method of dynamic value selection to select the resolution coefficient. Let $\Delta_v(k) = ([\Delta_k]/m)$, then $\varepsilon(k) = (\Delta_v(k)/\ \max_i \max_k [\Delta_k])$, $\rho = \{\rho(1), \rho(2), \ldots, p(n)\}$. The value of $\rho(k)$ is as follows: when $(1/\varepsilon(k)) > 3$, take $\rho(k) = 1.5\varepsilon(k)$; when $2 \le (1/\varepsilon(k)) \le 3$, take $\rho(k) = 2\varepsilon(k)$; when $0 < (1/\varepsilon(k)) < 2$, $\rho(k) = 0.9$; when $\varepsilon(k) = 0$, and $\rho(k) =$ arbitrary value.

(4) Compare the sample indicators to be evaluated and the indicators in the standard sample one by one. Then, calculate the correlation coefficient ξ_{ij} and the indicator weight ω_j according to formula (3). Then, the comprehensive gray relational degree of the sample is evaluated and compared with the ideal sample gray relational degree as shown in the following equation:

$$\Phi_i = \sum_{j=1}^{n} \omega_j \xi_i(j), \quad \Phi_i \in [0, 1], \tag{4}$$

where $i = 2, 3, \ldots, m$; $j = 1, 2, \ldots, n$; and ω_j is the weight of the j-th index.

4.2.2. The Entropy Method. The process of using the entropy method to determine the objective weight of the index is as follows: in an evaluation problem with n evaluation indexes and m evaluation levels, the entropy value of the i-th individual evaluation index is defined as the following equation [24]:

$$H_i = -\frac{1}{\ln m} \sum_{j=1}^{m} f_{ij} \ln f_{ij}, \quad i = 1, 2, \ldots, n, \tag{5}$$

where $f_{ij} \ne 0$, when $f_{ij} = 0$, $f_{ij} \ln f_{ij} = 0$.

The entropy weight of the i-th evaluation index is defined in the following equation:

Table 1: Retailer's MCIQ into 8 dimensions.

Index	Meaning
Process consistency	It refers to the consistency of process attributes compared between different channels, such as service perception, image, waiting time, and customer service level.
Information consistency	It refers to the consistency of response information obtained by consumers through different channels.
Emotional value	It refers to the measurement of shopping emotion. For example, online store shopping is exciting.
Procedural value	It refers to measuring the purchase process ease. For example, it is more convenient to shop in the online retailer store.
Transparency of service structure	It refers to the customer's perception of the attributes of all available channels. It will directly affect the choice of channels for customers to receive specific services. The transparency of the service structure reflects the level of the retailer's multiple channel service structure, the consistency of information, and business relevance. Process consistency represents the three forms of interaction and integration between channels.
Online result value	It refers to the measurement of the final result of online purchases, such as the value of the retailer's online store's goods and services.
Business relevance	It refers to the connection between interaction generated through a specific channel and interaction generated through other channels.
Online purchase intention	It is the willingness of consumers to buy online.

$$\omega_i = \frac{1 - H_i}{n - \sum_{i=1}^{n} H_i}, \tag{6}$$

meeting the conditions $\sum_{i=1}^{n} \omega_i = 1$, $0 \le \omega_i \le 1$.

The calculated $W = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is the objective weight value of each index determined by the entropy method.

4.3. Building the NARX Model. The NARX neural network, as a dynamic recurrent neural network with output feedback connection, can effectively overcome the phenomenon of error accumulation in time series data. Therefore, this article uses a nonlinear autoregressive dynamic neural network model for processing. This model belongs to a commonly used method in time series analysis [25]. It uses the combination of variables at several previous moments to infer the development of variables at subsequent moments. The process of NARX neural network model building is shown in Figure 2.

A typical NARX neural network comprises an input layer, hidden layer, output layer, and input delay function. Its basic structure is shown in Figure 3. In the figure, $X(k)$ represents the neural network's input value and $Y(k)$ represents the output value of the neural network.

The difference between a NARX neural network and the ordinary neural network adds a delay function before the hidden layer. The effect of the delay function on the output is as follows [26]:

$$y(t) = f(x(t-1), \ldots, x(t-d), y(t-1), \ldots, y(t-d)), \tag{7}$$

where d represents the delay order.

In the NARX neural network, the output signal is delayed and then input into the neural network, and the hidden layer and the output layer are combined to obtain the final output result [27]. i represents the number of input data, l is the number of hidden layer neurons, x_i represents the i-th input signal of the network, w_{ij} represents the connection weight between the i-th output delay signal and

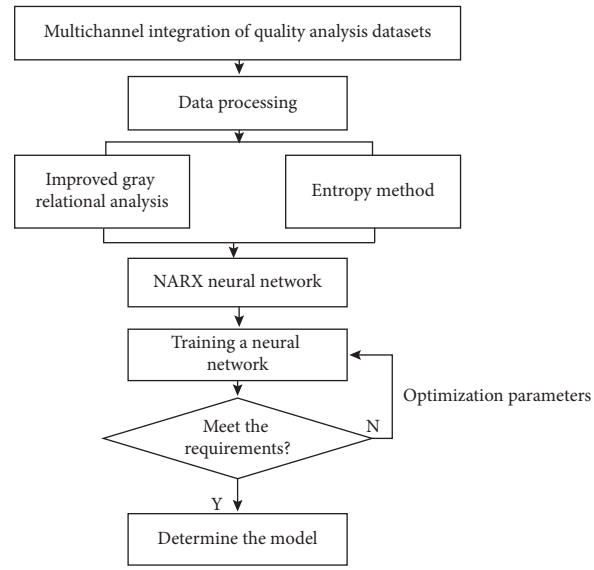


Figure 2: Flowchart of NARX neural network model setting.

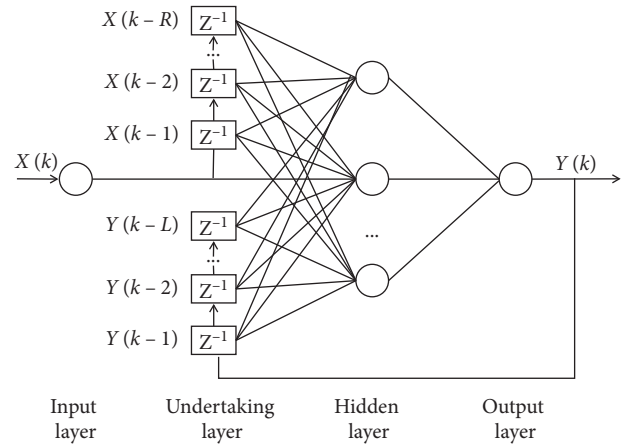


Figure 3: Schematic diagram of NARX neural network structure.

the j-th neuron, a_j represents the threshold value of the j-th hidden layer neuron, and the hidden layer activation function f is integrated to obtain the calculation result of each neuron as shown in the following equation [28]:

$$H_j = f\left(\sum_{i=1}^{n} w_{ij}x_i + a_j\right). \tag{8}$$

Using w_{ij} as the connection weight between the j-th neuron in the hidden layer and the output layer neuron [29] and b as the output layer neuron threshold, calculate the output as follows:

$$O = f\left(\sum_{j=1}^{l} H_j w_j + b\right). \tag{9}$$

5. Multiple Channel Integration Quality Assessment

5.1. Evaluation of Data Collection. According to the comprehensive evaluation index system of multisource channel integration quality determined above, fifty thousand consecutive samples of a commodity in 3 months are used as a dataset. The data fields in the dataset are the consistency of the collection process x_1, information consistency x_2, emotional value x_3, procedural value x_4, service structure transparency x_5, online result value x_6, business relevance x_7, and online purchase intention x_8. Table 2 shows sample records in the evaluation datasheet.

5.2. Evaluation Using Improved GRA Method. The GRA method is widely used, but there are still some defects. The resolution coefficient ρ is usually set to 0.5 based on experience or uniformly set to 0.5. This paper adopts the method of determining the resolution coefficient's dynamic state to improve the GRA method. According to the data series's stability, this paper adopts the method of dynamic value selection to select the resolution coefficient. Fifty thousand consecutive samples of a product for three months are selected as the dataset to verify the method. Compared with the traditional gray correlation method, the improved gray relational degree suppresses outliers' influence in the observation sequence on the correlation space, making the correlation analysis more realistic. Figure 4 shows the weight of the evaluation index.

5.3. Comprehensive Evaluation Using NARX Neural Network. This NARX model takes process consistency x_1, information consistency x_2, emotional value x_3, procedural value x_4, service structure transparency x_5, online result value x_6, business relevance x_7, and online purchase intention x_8 as input. The comprehensive improvement in the gray relational degree x_9 is the output of the NARX model. The input data are inputted into the NARX neural network for training. The network training method uses the Levenberg–Marquardt algorithm, the number of hidden layer neurons is set to 10, and the delay is set to 2. It can be seen from Table 3 and Figure 5 that the network has an excellent fitting effect, so the training of this model is ideal.

Table 2: Evaluation data.

Sample number	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	3.0	1.0	1.0	3.0	1.0	5.0	489.0	0.0417
2	1.0	−1.0	−1.0	4.0	1.0	4.0	30.0	0.0376
3	1.0	2.0	1.0	3.0	0.0	1.0	545.0	0.0000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50000	−1.0	1.0	1.0	2.0	0	4.0	244.0	0.0000

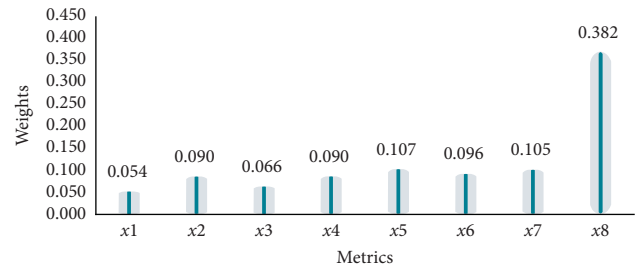


Figure 4: Evaluation index weight diagram.

Table 3: Training results.

Result	Target	MSE/10^{-7}	R/10^{-1}
Train	35000	1.53876	9.96358
Verify	15000	1.69412	9.95800
Test	15000	1.56598	9.96437

5.4. Comparison of Evaluation Methods. The MSE of the GRA method is 0.0000153, and the MSE of the NARX neural network method is 0.000000180. It can be seen from Figure 6 that the output value produced by the NARX neural network is closer to the target output value, while the output error of the GRA method is more massive. Therefore, NARX neural network comprehensive evaluation is more applicable, more accurate, and can more effectively reflect isolated networks' autonomous operation capability.

5.5. Multiple Channel Integration Quality Capability Score. The evaluation results with standardized scores between [1, 0.75) are regarded as excellent, [0.75, 0.5) as medium, [0.5, 0.25) as qualified, and [0.25, 0) as unqualified. Selecting a group of typical samples as the reference samples for MCIQ capability evaluation, the evaluation results are shown in Table 4.

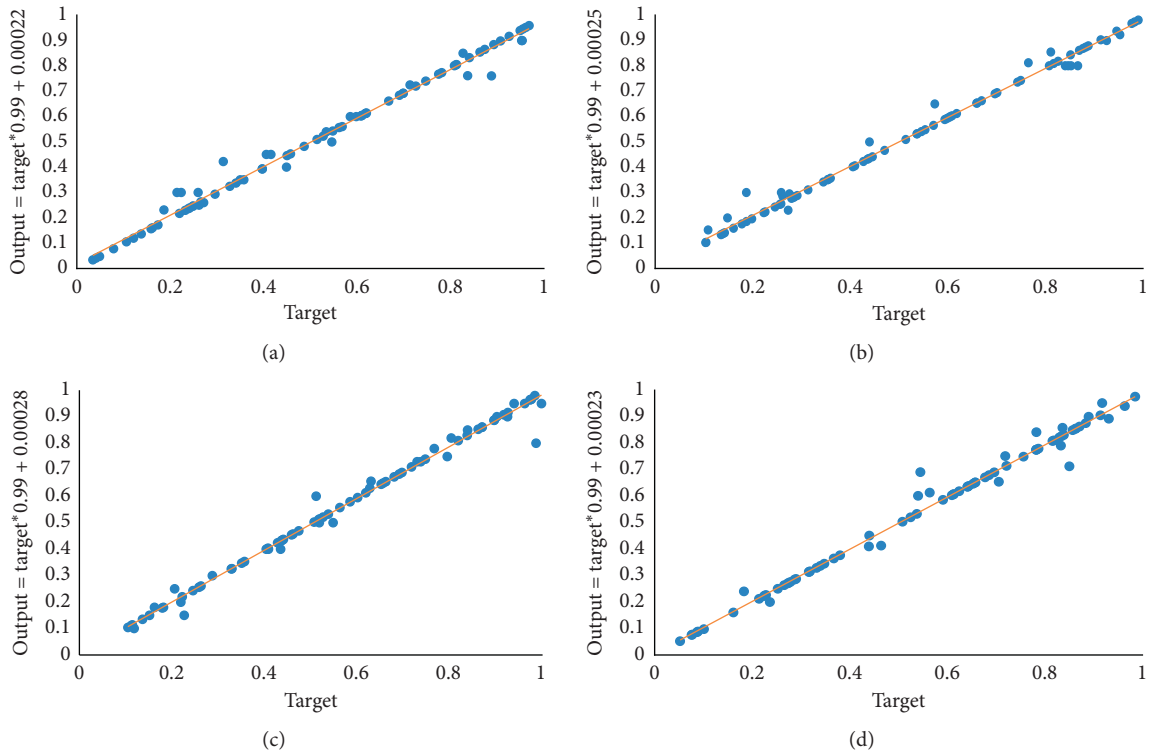


Figure 5: Neural network model fitting diagram: (a) $R = 0.9959$, (b) $R = 0.99435$, (c) $R = 0.99464$, and (d) $R = 0.99518$.

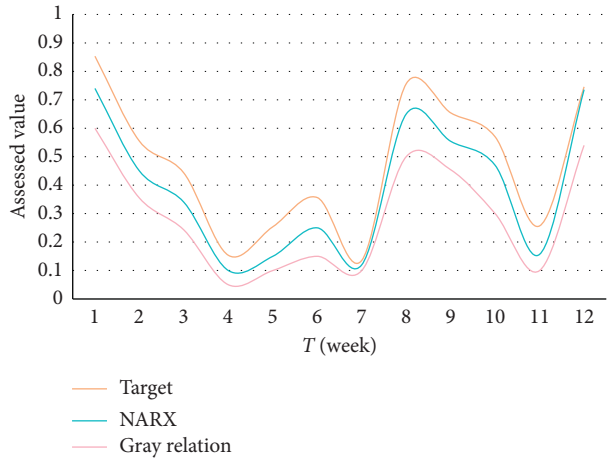


Figure 6: Comparison of evaluation results and target values.

Table 4: Evaluation results of typical samples.

Reference sample sequence	t (week)	Normalized score	Evaluation result
1	1	0.8532	Excellent
2	3	0.5547	Medium
3	6	0.4426	Qualified
4	4	0.1536	Unqualified

6. Conclusions and Future Work

This paper proposes a method for a comprehensive evaluation of MCIQ capability using the NARX neural network. This method's MCIQ index system comprehensively considers the impact of key factors affecting online purchase intentions based on analyzing multiple channel integration. The entropy method and the advantages of the improved

GRA method are used in the data processing. Then, NARX neural network modeling is used to avoid the disadvantages of the GRA algorithm. The evaluation results of multiple scenarios verify that this method is effective and applicable.

Although this study has made a certain breakthrough in the quality assessment of multiple channel integration, the results still have some limitations. This research has not considered shoppers' characteristics, such as shopping orientation, involvement, and time pressure. Future research can further explore the influence of these factors on the model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Berman and S. Thelen, "A guide to developing and managing a well-integrated multi-channel retail strategy," *International Journal of Retail & Distribution Management*, vol. 32, no. 3, pp. 147–156, 2004.

[2] K. A. Saeed, V. Grover, and Y. Hwang, "Creating synergy with a clicks and mortar approach," *Communications of the ACM*, vol. 46, no. 12, pp. 206–212, 2003.

[3] C. M. Chan and S. L. Pan, *Intertwining Offline and Online Channels in Multiple Channel Public Service Delivery: A Case Study*, Academy of Management, Briarcliff Manor, NY, USA, 2005.

[4] H.-H. Lee and J. Kim, "Investigating dimensionality of multichannel retailer's cross-channel integration practices and effectiveness: shopping orientation and loyalty intention," *Journal of Marketing Channels*, vol. 17, no. 4, pp. 281–312, 2010.

[5] L.-B. Oh and H.-H. Teo, "Consumer value co-creation in a hybrid commerce service-delivery system," *International Journal of Electronic Commerce*, vol. 14, no. 3, pp. 35–62, 2010.

[6] R. Sousa and C. A. Voss, "Service quality in multichannel services employing virtual channels," *Journal of Service Research*, vol. 8, no. 4, pp. 356–371, 2006.

[7] R. Madaleno, H. Wilson, and R. Palmer, "Determinants of customer satisfaction in a multi-channel B2B environment," *Total Quality Management & Business Excellence*, vol. 18, no. 8, pp. 915–925, 2007.

[8] T. M. T. Hossain, S. Akter, U. Kattiyapornpong, and Y. K. Dwivedi, "Multichannel integration quality: a systematic review and agenda for future research," *Journal of Retailing and Consumer Services*, vol. 49, pp. 154–163, 2019.

[9] L. M. Wang, Z. Y. Hao, X. M. Han, and R. H. Zhou, "Gravity theory-based affinity propagation clustering algorithm and its applications," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 4, pp. 1125–1135, 2018.

[10] J. Niresh, N. Archana, and G. Anand Raj, "Optimisation of linear passive suspension system using MOPSO and design of predictive tool with artificial neural network," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 105–110, 2019.

[11] J. Li, S. X. Pan, L. Huang, and X. Zhu, "A machine learning based method for customer behavior prediction," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1670–1676, 2019.

[12] D.-V. Căiman and T.-L. Dragomir, "Empirical voltage-current signatures for individual household consumers obtained by non-linear regression," *Studies in Informatics and Control*, vol. 28, no. 2, pp. 201–212, 2019.

[13] Y. Ge and H. Wu, "Prediction of corn price fluctuation based on multiple linear regression analysis model under big data," *Studies in Informatics and Control*, vol. 28, 2019.

[14] C. Ciurea and F. G. Filip, "Virtual exhibitions in cultural institutions: useful applications of informatics in a knowledge-based society," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 55–64, 2019.

[15] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation and Economic Cybernetics Studies And Research*, vol. 53, pp. 77–92, 2019.

[16] X.-L. Shen, Y.-J. Li, Y. Sun, and N. Wang, "Channel integration quality, perceived fluency and omnichannel service usage: the moderating roles of internal and external usage experience," *Decision Support Systems*, vol. 109, pp. 61–73, 2018.

[17] C. Zehir and E. Narcıkara, "E-service quality and e-recovery service quality: effects on value perceptions and loyalty intentions," *Procedia-Social and Behavioral Sciences*, vol. 229, no. 427, p. 43, 2016.

[18] J. Atanasijevic and D. Milosevic, "Upgrading the business intelligence system by implementing the decision tree model in the R software package," *Studies in Informatics and Control*, vol. 29, no. 2, pp. 243–254, 2020.

[19] I. Vrecko, J. Kovac, B. Rupnik, and B. Gajsek, "Using queuing simulation model in production process innovations," *International Journal of Simulation Modelling*, vol. 18, no. 1, pp. 47–58, 2019.

[20] M. S. Yang, L. Ba, Y. Liu et al., "An improved genetic simulated annealing algorithm for stochastic two-sided assembly line balancing problem," *International Journal of Simulation Modelling*, vol. 18, no. 1, pp. 175–186, 2019.

[21] L. Lei, W. Chen, Y. Xue, and W. Liu, "A comprehensive evaluation method for indoor air quality of buildings based on rough sets and a wavelet neural network," *Building and Environment*, vol. 162, 2019.

[22] C.-T. Su and F.-F. Wang, "Integrated fuzzy-connective-based aggregation network with real-valued genetic algorithm for quality of life evaluation," *Neural Computing and Applications*, vol. 21, no. 8, pp. 2127–2135, 2012.

[23] J. Zhao, G. Ji, Y. Tian, Y. Chen, and Z. Wang, "Environmental vulnerability assessment for mainland China based on entropy method," *Ecological Indicators*, vol. 91, pp. 410–422, 2018.

[24] H.-Y. Wu, a. Tsai, and H.-S. Wu, "A hybrid multi-criteria decision analysis approach for environmental performance evaluation: an example of the TFT-LCD manufacturers in Taiwan," *Environmental Engineering and Management Journal*, vol. 18, no. 3, pp. 597–616, 2019.

[25] J. Li, U. Konuş, F. Langerak, and M. C. D. P. Weggeman, "Customer channel migration and firm choice: the effects of cross-channel competition," *International Journal of Electronic Commerce*, vol. 21, no. 1, pp. 8–42, 2017.

[26] B. Xue, M. Liu, and Q. Sun, "Applications of grey relational analysis to enterprise performance evaluation of express listed companies in China," in *Proceedings of the 2017 3rd*

International Forum on Energy, Environment Science and Materials (IFEESM 2017), Shenzhen, China, November 2017.

[27] A. Coser, M. M. Maer-Matei, and C. Albu, "Predictive models for loan default risk assessment," *Economic Computation And Economic Cybernetics Studies And Research*, vol. 53, pp. 149–165, 2019.

[28] E. Comăniță, P. Cozma, I. Simion, M. Roşca, and M. Gavrilescu, "Evaluation of eco-efficiency by multicriteria decision analysis. case study of eco-innovated and eco-designed products from recyclable waste," *Environmental Engineering and Management Journal*, vol. 17, pp. 1791–1804, 2018.

[29] P. P. Dey, S. Pramanik, and B. C. Giri, "An extended grey relational analysis based multiple attribute decision making in interval neutrosophic uncertain linguistic setting," *Neutrosophic Sets and Systems*, vol. 11, pp. 21–30, 2016.

Hindawi
Complexity
Volume 2020, Article ID 5639787, 14 pages
https://doi.org/10.1155/2020/5639787

Research Article

Towards a Framework for Acquisition and Analysis of Speeches to Identify Suspicious Contents through Machine Learning

Md. Rashadur Rahman,[1] Mohammad Shamsul Arefin,[1] Md. Billal Hossain,[1] Mohammad Ashfak Habib,[1] and A. S. M. Kayes[2]

[1]*Department of Computer Science & Engineering, Chittagong University of Engineering & Technology, Chattogram, Bangladesh*
[2]*Department of Computer Science & Information Technology, La Trobe University, Melbourne, Australia*

Correspondence should be addressed to A. S. M. Kayes; a.kayes@latrobe.edu.au

Received 13 July 2020; Revised 19 September 2020; Accepted 29 October 2020; Published 16 November 2020

Academic Editor: Abd E.I.-Baset Hassanien

The most prominent form of human communication and interaction is speech. It plays an indispensable role for expressing emotions, motivating, guiding, and cheering. An ill-intentioned speech can mislead people, societies, and even a nation. A misguided speech can trigger social controversy and can result in violent activities. Every day, there are a lot of speeches being delivered around the world, which are quite impractical to inspect manually. In order to prevent any vicious action resulting from any misguided speech, the development of an automatic system that can efficiently detect suspicious speech has become imperative. In this study, we have presented a framework for acquisition of speech along with the location of the speaker, converting the speeches into texts and, finally, we have proposed a system based on long short-term memory (LSTM) which is a variant of recurrent neural network (RNN) to classify speeches into suspicious and nonsuspicious. We have considered speeches of Bangla language and developed our own dataset that contains about 5000 suspicious and nonsuspicious samples for training and validating our model. A comparative analysis of accuracy among other machine learning algorithms such as logistic regression, SVM, KNN, Naive Bayes, and decision tree is performed in order to evaluate the effectiveness of the system. The experimental results show that our proposed deep learning-based model provides the highest accuracy compared to other algorithms.

1. Introduction

Speech has been the mostly used medium for conveying information among people all over the world since the dawn of civilization. Speech is the most effective method of addressing and communicating with the audience in order to deliver some message. Speech empowers an individual to reach a large number of people directly. It is a very dynamic way to shift a huge number of people's mindset or to reinforce their confidence in speaker [1]. Historically, it played a significant role in persuading the audience into a specific agenda [2, 3].

Speech can be used to influence people in both righteous and wrong ways. Speech has been used to escalate hatred among the communities [4–6]. Misguided speech leads people in the wrong direction, which raises social risk. The preservation of the right to freedom of speech is an integral aspect of modern democratic states [7]. People are freely expressing their emotion, thought, anger, and grudge through speeches. Often this freedom of expression is misused by certain people in society, which causes social controversies [8, 9]. The problem is more critical when the religious people give deceptive speeches. This is because, in general, people have love for their religions and in most of the cases they respect and rely on the speeches of their religious speakers [10]. Wrong speeches can contribute to crime and pose a threat to the government. Since Internet technology is growing rapidly, any misleading speech can be easily spread among different groups of people through various social media platforms [11]. Statistics show an alarming rate of growth of hate crime over the years[2]. It not only threatens a country's people's lives and livelihood but also undermines

worldwide peacekeeping. These types of threats are troubling in both national and international security and steps are being taken to avoid these kinds of crimes.

To prevent any potential crime resulting from speech, suspicious speeches must be identified in the shortest time possible. There are a lot of speeches being delivered daily, which are very difficult for manual inspection. Therefore, a good speech repository and speech monitoring system can be very useful in spreading good speeches and restraining suspicious speeches from spreading. If the suspicious speeches can be classified beforehand, it will be very handy for law-enforcement agencies to take proactive steps to deter any unwanted incidents. Moreover, if the speech audio can be converted into text, then the speech can be automatically interpreted to extract information from the converted text. Analyzing the speech from direct voice data is quite difficult due to noises in speech recording and varieties of speeches. Audio data analysis is a very complex task, which includes multidimensional analysis as a simple audio can have millions of different data segments. Many attributes like frequency, pitch, volume, dialect, and so forth must be considered in the analysis of audio data. Therefore, analysis of direct audio speech is quite complex and requires much more processing time; most importantly, it requires much computational power, which may be impractical for devices with limited processing ability like smartphones. So, converting the speeches into text before analysis and then analyzing the converted text are more convenient and time-efficient.

The usage of mobile phones has increased drastically over the last few years. It is approximately 3.5 times larger than PCs [12]. Nowadays, mobile phones not only are being used as a tool for making calls and writing SMSs but also act as a means for personal entertainment and communication with the world [13]. Almost every feature that is available in a PC can also be found in a smartphone. Smartphones are available to almost everyone and one of the most popular operating systems being used in those devices is Android, developed by Google [14].

Classification of speech refers to the task of classifying a speech into a set of predefined classes. Classifying speeches into suspicious and nonsuspicious is very necessary for reducing virtual social harassment, predicting criminal activities, social clashes, and riots, and ensuring overall national security. Such a system has not yet been built to detect suspicious Bangla speech. Here we proposed a framework for acquisition of speeches and a deep learning method which is based on LSTM to detect suspicious speeches. To the best of our knowledge, this is the first work to detect suspicious Bangla speech. The contributions of our work are summarized as follows:

(i) We develop a mobile application for the acquisition of speech efficiently along with the location of the speaker. The application stores the speeches, detects the language of the speech, and finally converts the speech into text using a speech recognition API.

(ii) We develop our own dataset for training and testing the models. The dataset contains about 5000 suspicious and nonsuspicious samples.

(iii) We propose a model based on LSTM for classifying the texts (converted from speech) into suspicious and nonsuspicious. We compare the accuracy of the model with other machine learning algorithms.

The rest of the paper is organized as follows. Section 2 includes a short overview of similar research works. The architecture of our proposed methodology is presented in Section 3. Section 4 provides a description of our dataset preparation. The implementation along with the evaluation of the system is shown in Section 5. Finally, Section 6 includes the conclusion and future research.

2. Related Works

The computational study of suspicious speech or hate speech detection, from computer science point of view, is in early phase. As we first convert our audio speeches into texts then classify the texts, this work falls into the domain of text classification. Machine learning has gained much more attention of the researchers in automatic detection of suspicious texts [15]. In machine learning, identification of suspicious speech is considered as classification problem. The vast majority of the studies found this as a problem of binary classification (suspicious speech versus nonsuspicious speech) [16].

Nobata et al. proposed a supervised classification method to detect abusive English comments [17]. They used their custom-build corpus for training and testing, which was developed by extracting comments on Yahoo! Finance and News. They used regression model for classification by analyzing different aspects of user comments. They divided the texts features into four classes: N-grams, Linguistic, Syntactic, and Distributional Semantics. For N-grams features, they used space included character N-grams (3 to 5 characters) and token unigrams and bigrams. Different combinations of features were used for classification for achieving notable accuracy. Vidgen and Yasseri developed an automated software tool to distinguish between strong-Islamophobic, weak-Islamophobic, and non-Islamophobic tweets [18]. They used 4000 annotated tweets as training set and used a combined feature selection model. Their SVM-based classifier obtained 77.6% accuracy. Oriola and Kotze [19] developed an English corpus of South African tweets and applied various machine learning algorithms to detect offensive speech. Their optimized SVM with character N-gram performed best with true positive rate of 0.894.

Deep learning is also applied to classify large amount of texts with notable accuracies. For classification of texts, deep learning methods enable the deep neural networks (DNN) by using their multiple stacked layers to learn abstract feature representation from input data. Most of the works in this domain use one-hot encoding based on word/characters as input features to their models [20, 21]. Some of the works combined multiple methods to classify texts. Zhang et al.

proposed a deep neural network model combining convolutional neural network (CNN) and gated recurrent unit (GRU) networks for detecting hate speech in Twitter [22]. Elastic net regularization was used along with optimized dropout and pooling layers. Word embedding was used for mapping texts into vectors. The output of the embedding layer was fed to one-dimensional convolutional layer for feature extraction. The extracted features were given to the GRU layer. They used publicly available twitter data. In [23], Risch et al. present various deep learning approaches for sentiment analysis in online platforms for detecting toxic comments. They propose fine-grained classification instead of binary classification. Salminen et al. considered four platforms, YouTube, Wikipedia, Twitter, and Reddit, and collected 197,566 comments and labeled these comments as hateful and nonhateful [24]. They experimented with several classification algorithms. They found that XGBoost performs better than others and BERT features are most impactful.

A number of researches have been carried out in the field of text classification of English texts. No significant research has yet been done in the Bangla text classification. For sentiment mining of Bangla text, Taher et al. proposed a system based on support vector machine (SVM) [25]. In their work, they applied both linear and nonlinear SVM to determine whether it is positive or negative sentiment. They generated their dataset from the comments of several Bangla online news sites. For the preprocessing of the text, they only considered adverb, adjective, selected nouns, and verbs. They reduced the verbs into their base forms to reduce the number of vectors. For vectorization, they applied N-gram method, where $N = 1$, 2, or 3. In terms of preserving sequence of the words (syntactic and semantic content), N-gram model is better than bag-of-words model for feature representation. However, within a sentence, related words can have a high distance, which may lead to misinterpretation of the context. In [26], Chy et al. applied Naive Bayes classifier for classifying web crawled Bangla news documents. They preporcessed the text by applying stemming, removal of the less significant words called stop-words, and single-letter words. For selecting features, they used inverse document frequency (IDF) method. Dhar et al. proposed method based on Multinominal Naive Bayes (MNB) classifier to classify Bangla documents into eight predefined classes [27]. For feature extraction and selection, they applied inverse class frequency (ICF) along with the term frequency- (TF-) inverse document frequency (IDF) feature selection method named as TF-IDF-ICF scheme. A comparison with other schemes TF-IDF and TF was also shown.

Sharif et al. proposed a system based on logistic regression for classifying suspicious Bangla text [28]. As there is no available dataset on suspicious Bangla text, they developed their own private corpus of suspicious Bangla text. Their proposed model was trained with only 1500 samples and the model was tested with 500 samples. Finally, they showed that logistic regression performed better in terms of accuracy (92%) among those algorithms. They used word frequencies as the feature of the model. Bag-of-words model is used for representation of features. The main limitation of bag-of-words approach is that it only counts the frequencies of the words but the sequence of the words is ignored. So, it can contribute to misclassification if the words are used to represent different contexts [16]. In [29], Ishmam et al. developed a dataset of 5,126 comments from public Facebook groups and classified them into six classes. Their gated recurrent unit (GRU) based model achieved at most 70.10% accuracy in detecting Bangla hate language. They have not clearly defined the six classes and a smaller number of training samples in each class yields poor accuracy. The RNN based model proposed in [30] achieved 82.20% accuracy on their collected 4,700 Bangla text samples from various online platforms. Islam et al. proposed a classification method based on Multinomial Naive Bayes (MNB) for spam detection of Bangla texts [31]. They only collected 1,965 instances from social media platforms like Facebook and YouTube. The model has an accuracy of 88.44%. To detect abusive languages and threats on social media, Chakraborty et al. showed a methodology using SVM with linear kernel, which obtained 78% accuracy [32]. Their dataset contains 5,644 comments and posts from different Facebook posts.

3. Methodology

In this section, we detail the overall architecture of our proposed system. As the input of our system is speech, we first convert the speech into corresponding texts before further analysis. First, we elaborate the speech acquisition and conversion into texts. Then we explain our LSTM based classification model in detail. Our overall framework consists of four modules: (i) speech acquisition module, (ii) speech storage module, (iii) speech recognition module, and (iv) speech analysis module. The overall graphical structure of the system is shown in Figure 1.

3.1. Speech Acquisition. Acquisition of speech includes two parts: one for speech recording and another for location tracking. The architecture of the speech acquisition module is shown in Figure 2. The speaker's sound is recorded by a microphone and stored in local storage as an audio file. For tracking the location of the speaker, latitude and longitude are calculated first. Making use of latitude and longitude, we determine the speaker's actual location. When the speaker finishes his speech, it is uploaded into the cloud database where further processing takes place.

3.2. Speech Storage. After acquisition of speech, it is stored. For each incoming audio file, a row is created in the database with a unique identifier. The file name of the incoming audio file may be the same as some other audio files in the database. So, the file name is renamed as follows: Unique ID + "." + File Extension. After analyzing the audio file, the converted text file of each audio is stored along with the speech location and language.

3.3. Speech Recognition. The language of the incoming speech is identified in the recognition module and the voice is translated to text. The overall architecture of the speech

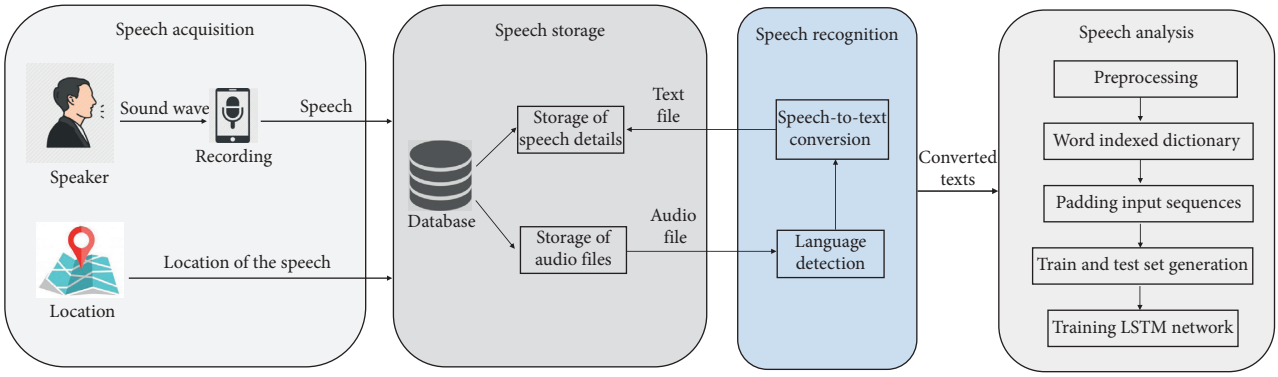


Figure 1: Overall graphical structure of the system.

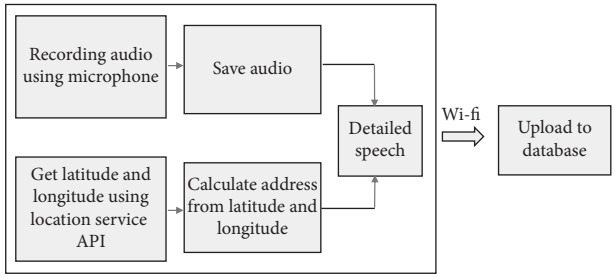


Figure 2: Architecture of speech acquisition module.

conversion process is shown in Figure 3. For converting the long-duration speech accurately, there are two possible ways. We can either split the speech into smaller chunks of constant size or split it based on the silence presented in the audio. If we split it by keeping a constant duration, then a word within the speech might get split and that word will not be detected. So, we choose to split the speech based on the silence presented. It is possible because when we speak, we pause for a small duration after finishing a sentence. That means we can consider each chunk as a sentence. The speech is split if the amplitude level in the audio is less than -16 dBFS for more than 0.5 seconds.

Real-life speeches can be in different languages. So, it is very important for your system to detect the language of the speech. Most of the time, speech contains more than one language and the system will not give good result for such type of mixed-language speeches. We need to specify the language to which we want to convert it. For example, if a speech contains Bengali language, then we need to specify language parameter as Bengali in the API. So, if we want to recognize speech containing multiple languages, we need to specify the language parameter manually. Our speech recognition module can detect three languages: Bangla, English, and Arabic. For this, we send two requests in the API for three different languages for each sentence or chunk. Then the converted text is split into words and then checked in the dictionary of that language if that word exists. If the word is found in the dictionary, then the counter for that language is increased. The language that has the maximum counter value is chosen as the language of that sentence. Algorithm 1 shows the process of detecting language and conversion of speech chunks into text.

Our system is developed to detect and convert speeches of three languages: Bangla, English, and Arabic. However, the analysis of the speech is performed only on Bangla speeches.

3.4. Speech Analysis. After converting the speeches into corresponding texts, our speech dataset has become a dataset of texts. This section describes the analysis of the converted texts from the speeches.

3.4.1. Preprocessing. Preprocessing of data is a vital part for training any machine learning model. As we train our model by text documents, the first step of preprocessing is tokenization. Tokenization is the process of splitting a text into group of streams of characters called tokens delimited by white space, new line, tab, and so on. After tokenization, each text document is a list of words (tokens). As all the words are not equally important in determining the context of the text, some words are removed to increase the accuracy of the model and to reduce the feature dimension. So, punctuation marks, English and Bangla numerals, special characters, least frequent words, and most frequent words are removed as they represent no significance to the context of the text.

3.4.2. Word Indexed Dictionary. Texts cannot be given as the direct input to the neural networks because neural networks do not accept direct text data as input. Neural networks only accept numeral inputs. As text is a sequence of words, if each word has an integer representation, then we can convert a sequence of words into a sequence of numbers, which can be fed to embedding layer. Algorithm 2 demonstrates the process of creating a word indexed dictionary, where each unique word is mapped to an integer. The algorithm starts by setting the index to 1 followed by iterating each word of all the text documents and updates the Word_to_Index dictionary. Same word can be encountered more than once. At each iteration, a word is encountered and it is checked whether or not the word is in the dictionary. If the encountered word is not in the dictionary, the current index is

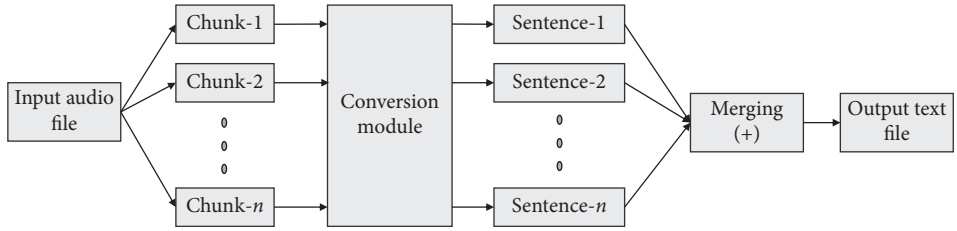


Figure 3: Framework architecture of the speech recognition module.

```
(1) Input: chunk
(2) Goal: language detection, converted text
(3) max ⟵ 0
(4) for i ⟵ 1 to 3 do
(5)     counter ⟵ 0
(6)     converted_text ⟵ recognition API (audio ⟵ chunk, language ⟵ i)
(7)     for each word in converted_text do
(8)       if word is in dictionary of language i then
(9)         counter ⟵ counter + 1
(10)      end if
(11)    end for
(12)    if counter > max then
(13)      max ⟵ counter
(14)      language ⟵ i
(15)      text ⟵ converted_text
(16)    end if
(17) end for
```

Algorithm 1: Detecting language and converting speech chunk into text.

assigned to that word and the index is increased; otherwise, the word is ignored.

Each text input is converted to a sequence of integers using the word indexed dictionary. Each word is represented by an integer.

3.4.3. Padding Input Sequence. The neural networks require inputs that have the same size and shape for both training and testing the network. Till now, we represented each piece of text data as a sequence of numbers. In our dataset, not all texts have the same number of words. So, the number sequences are of variable lengths, but the LSTM network takes input of the same length and dimension. So, we need to have the input sequences with the same size and that is why we need to pad the input sequences to the maximum length. The maximum length is set to 200. There are two types of padding: prepadding and postpadding. In prepadding, all input sequences, which are shorter than the maximum length sequence, are padded with zeros in the beginning. In case of postpadding, the sequences are padded with zeros in the ending. We used prepadding in our system as it is more suitable with LSTM [33].

3.4.4. Word Embedding. To represent the features of the words, we used word embedding technique. Word embedding is a form of word representation which allows words that are used in similar ways to have similar representations. In embedding, each individual word is represented as a real-valued vector in a predefined vector space. A word's position within the vector space is learned from text and is based on the words surrounding the word when used in the text. Each word is mapped to a real-valued vector of higher dimensions. We have used an embedding layer as the first hidden layer of our LSTM network in which word embedding is learned jointly with LSTM model. We defined our embedding layer with the size equal to the size of our vocabulary, a vector space of 200 dimensions in which words will be embedded.

3.4.5. Training and Test Set Generation. After converting the speeches into corresponding text documents, we considered the accuracy of the conversion for considering as a candidate in training or test set. The speeches which are converted with more than 90% accuracy are considered as the candidates of our training and testing set. Our training set $T = t_1, t_2, t_3, \ldots, t_n$ contains n number of text documents. Each of the text documents is labeled as either suspicious or nonsuspicious. The suspicious class is denoted as (C_S) and nonsuspicious class is denoted as (C_{NS}). Our test set also contains labeled text documents; this set is used to validate the model. Our dataset contains about 5000 samples of both suspicious and nonsuspicious speech. We considered 4000 speeches for training the model and 1000 ones for validating the model.

```
(1) Input: lists of tokenized documents, where Documents[i] represents the list of words present in i^th document.
(2) Goal: a word-to-index dictionary of input documents, Word_to_Index[w] gives the index of the word w
(3) index ⟵ 1
(4) Word_to_Index ⟵ {}
(5) for document_i in do
(6)       for word in document_i do
(7)          if word not in Word_to_Index then
(8)             Word_to_Index[word] ⟵ index
(9)             index ⟵ index + 1
(10)         else
(11)            do nothing (the word is already added to the dictionary)
(12)         end if
(13)      end for
(14) end for
```

Algorithm 2: Creating word indexed dictionary.

3.4.6. LSTM Network. Long short-term memory (LSTM) is a variant of RNN, a class of deep neural networks (DNN). RNNs emerged as efficient learners of sequential data. As text is sequence of word and preserving sequence is very important to interpret the actual context of the text, RNN based models are most suitable for text analysis [34]. RNNs are suitable for sequential data because, unlike other neural networks where all the inputs are independent from each other, in RNN inputs are interrelated.

Although RNNs are very powerful for learning sequence, they are practically vulnerable to the vanishing gradient problem [35]. Vanishing gradient problem means RNN fails to remember things in long past. It is possible that the sentiment of a text document can highly rely on the beginning portion of the text. So it can lead to misclassification of the text document as simple RNN cannot remember long-term dependencies. In order to handle the vanishing gradient problem, an improved variant of simple RNN is developed, which is called LSTM [36]. The structure of our proposed LSTM network is shown in Figure 4.

The length of the maximum sequence of words is set to 200. Zero padding is used for the text shorter than 200. Each text document is converted to a vector of integers $I_t = [w_1, w_2, \ldots, w_{200}]$, where each integer w_i represents a unique word. Our model takes vector of length 200 as input. We used an embedding layer of dimension 200. The embedding layer transforms each word (represented by integer) into a vector representation of length 200. The output of the embedding layer is then given to the LSTM layer. The LSTM layer contains 128 memory units. Each of the feature values is multiplied by the weights of each LSTM cell. The activation function "tanh" is used in the LSTM unit and the recurrent activation function is "sigmoid." The weighted sum of the dense layer of 128 units is used to map the output of the final output layer of single unit. We used "Binary Cross-Entropy" as the loss function and "Adam" optimizer to train the model [37].

3.4.7. Classification. Our proposed deep learning model based on LSTM is a binary classifier which classifies the input texts into suspicious (C_S) and nonsuspicious (C_{NS}) classes. More specifically, as it is binary classification problem, we have only one output neuron in our network. The output is given as the probability that the given text is suspicious. If the probability is greater than 0.5, then it is classified as suspicious (C_S); otherwise, it is classified as nonsuspicious (C_{NS}).

4. Dataset Preparation

The most important aspect of every experiment is the creation of a dataset. Dataset quality plays a prime role in any experiment's performance. The available amount of dataset in Bangla language is very low. In the domain of suspicious or hate speech detection, authors do not usually use publicly available datasets and they do not publish their own datasets [11]. There is no standard dataset available on Bangla suspicious text data, so we built our own dataset of suspicious and nonsuspicious Bangla text. Developing such dataset is a time-consuming and tedious task, as it demands a lot of attention; thus, defects cannot be introduced into the upcoming systems as long as sufficient data can be integrated into the dataset for robust program assessment.

One of the most challenging tasks of creating this dataset is to define what is suspicious speech. Nobata et al. define such speech as the language that attacks a group or community, which is based on religion, ethnic origin, gander, age, and disability [17]. There are many different definitions of suspicious speech or hate speech from different sources. Fortuna et al. identified four dimensions in which the comparison of these definitions can be made [16]. Several properties of the suspicious activity are defined by U.S. Department of Homeland Security[3]. To label a speech as suspicious, we must follow some criteria. We have set some properties for suspicious speech. If any speech meets one or more of these properties, the speech is labeled as suspicious. We sum up these properties in four main domains for collecting suspicious speech. These are the following:

(i) Religious humiliation: speech with intentional and malicious intent to offend religious feelings of any

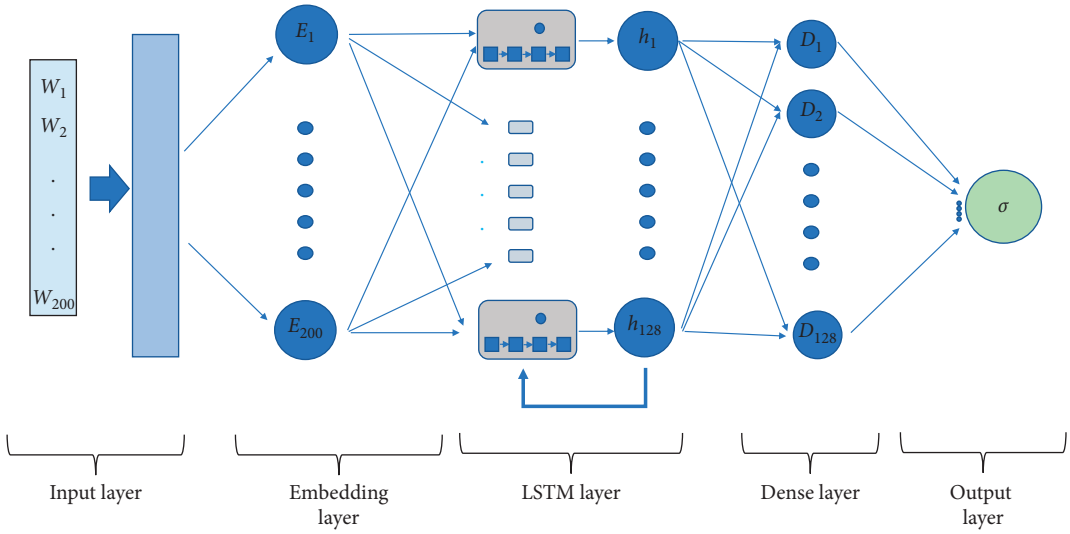


Figure 4: Proposed LSTM network.

class or speech aimed at insulting any religion or the religious beliefs.

(ii) Violence: speech that motivates people for any kind of terrorist activities, contains threat to the lives of people, or provokes violent activities.

(iii) Antigovernment: speech that provokes people against the government, law-enforcement agencies, or any community.

(iv) Offending nationalism: speech that disrespects the country or any national feeling.

The suspicious speeches are collected manually from YouTube, Facebook, and online blogs like Dhormockery 4, Shongshoy 5, and Istishon 6. The nonsuspicious speeches are also manually collected from various online resources.

4.1. Data Annotation. Three human annotators who are experts in the study of suspicious contents annotated the dataset. Based on our predefined four main domains of suspicious properties, all the 5,000 samples of data have been annotated blindly by these three experts. We calculated Fleiss' kappa [38], which is a statistical measure for determining the credibility of agreement between numbers of raters. The interrater agreement across the 5,000 samples was very satisfactory. The value of Fleiss' kappa was 0.959, which refers to almost-perfect agreement. There were still some disagreements between the annotators. In such case of disagreement, the speech was assigned to the class based on majority voting. The number of samples in each domain of the suspicious speech is shown in Figure 5.

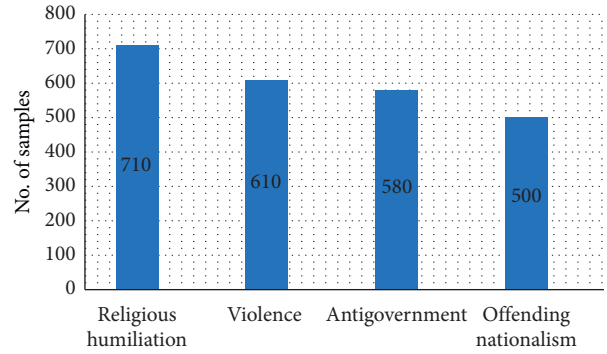


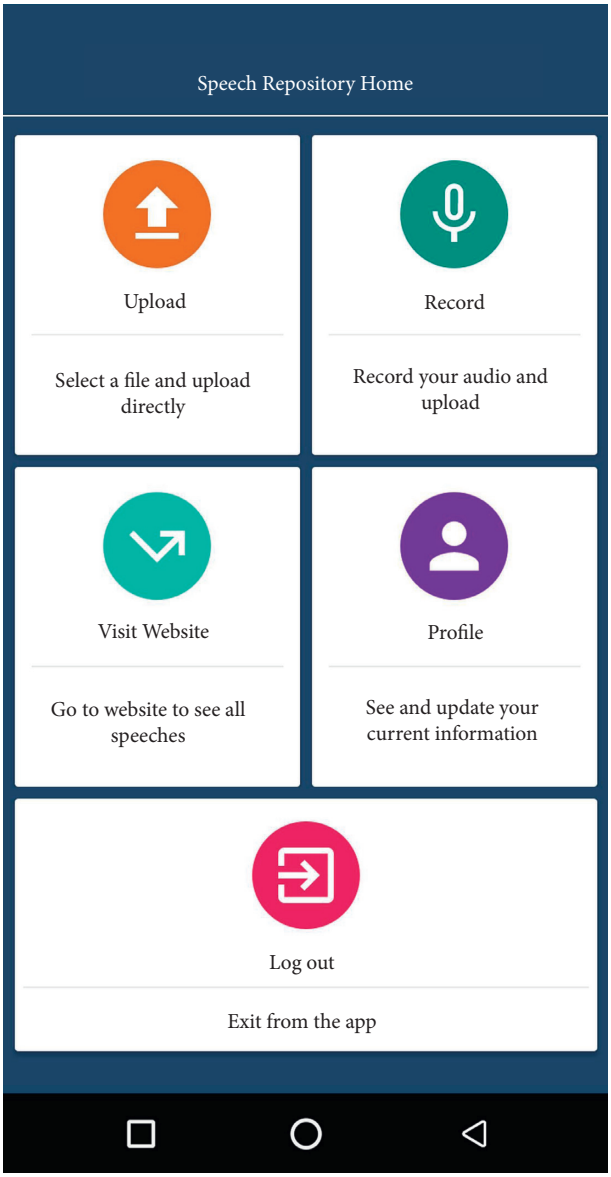
Figure 5: Number of samples in each domain of suspicious class.

5. Implementation and Evaluation

5.1. Implementation. For the purpose of acquisition of speech, we have developed an Android application. Acquisition of speech contains two parts: (i) recording of the speech and (ii) tracking the location of the speech. The phone's microphone is used for the purpose of recording the audio file. After recording, the file is stored in local storage. To track the location, the latitude and longitude are calculated. The user can also upload any audio file from the local storage. The interfaces of our Android application are shown in Figure 6.

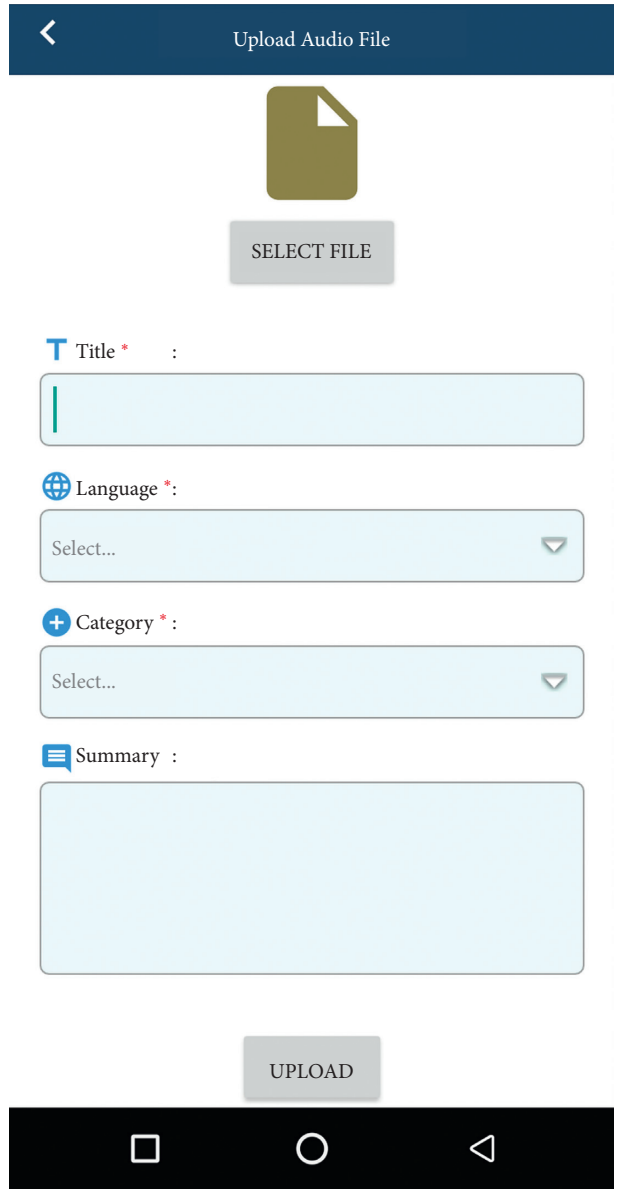
We used Google speech-to-text (gSTT) converter as our speech recognition API 7. From the latitude and longitude, the actual address of the speaker is calculated, which contains city, postal code, state, and country name. For validating the framework, we collected about 200 speeches from 10 different locations of Bangladesh from various speakers, both males and females. The performance of the speech acquisition process for 10 different locations is shown in Table 1. From the table, it can be seen that some location information was not available in some region (represented by null). It is because of the variation in geocoding detail. Moreover, our system can perform acquisition of speech with about 100% accuracy.

The performance of the speech recognition is dependent on the quality of the speech. The performance of our recognition module is shown in Table 2. From the table, we can see that speeches, which contain single language, have higher recognition accuracy than the speeches containing mixed languages. For a speech, if the total number of words is T,



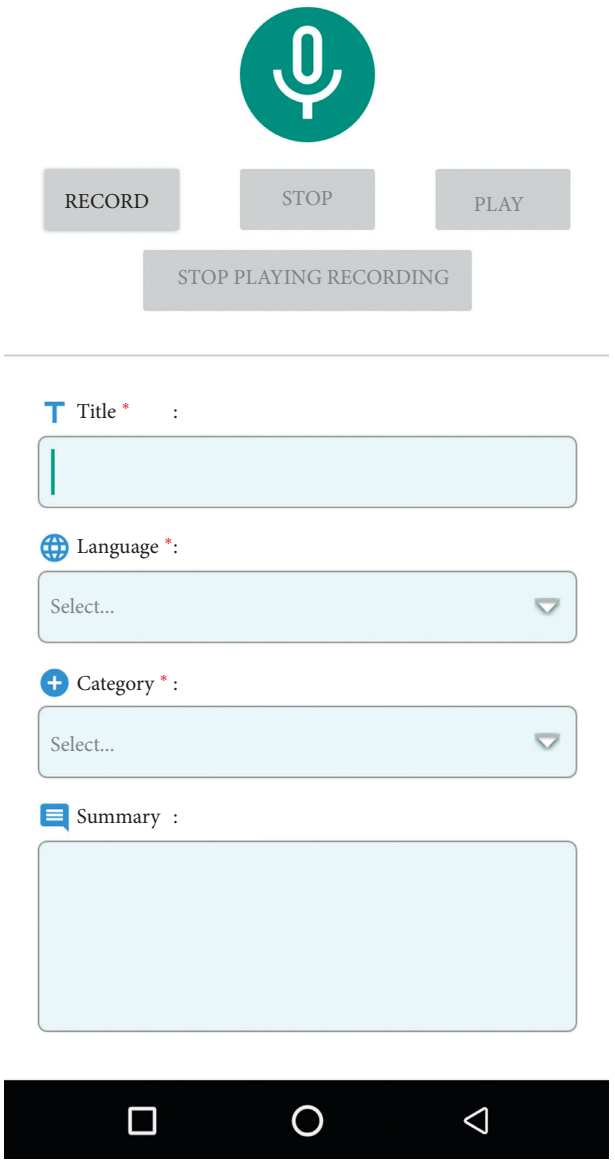
(a)

Figure 6: Continued.



(b)

Figure 6: Continued.



(c)

Figure 6: Interfaces of our Android application for acquisition of speech (a) Home interface, (b) Upload interface, (c) Record interface.

Table 1: Performance evaluation of speech acquisition process.

Speaker	Actual location	Detected location	Speech quality
1	Pahartoli, Chattogram, 4349, Bangladesh	Pahartoli, Chattogram, 4349, Bangladesh	Highly satisfactory
2	Habiganj, Sylhet, 3310, Bangladesh	Habiganj, Sylhet, 3310, Bangladesh	Highly satisfactory
3	Durgapur, Chandpur, 3640, Bangladesh	Durgapur, Chandpur, Null, Bangladesh	Satisfactory
4	Dhanmondi, Dhaka, 1208, Bangladesh	Dhanmondi, Dhaka, 1208, Bangladesh	Highly satisfactory
5	Gulshan, Dhaka, 1213, Bangladesh	Gulshan, Dhaka, 1213, Bangladesh	Highly satisfactory
6	Dinajpur, Rangpur, 5262, Bangladesh	Dinajpur, Rangpur, 5262, Bangladesh	Highly satisfactory
7	Jamalpur, Mymensingh, 2030, Bangladesh	Jamalpur, Mymensingh, 2030, Bangladesh	Highly satisfactory
8	Bogura, Rajshahi, 5892, Bangladesh	Bogura, Rajshahi, 5892, Bangladesh	Highly satisfactory
9	Barguna, Barisal, 8730, Bangladesh	Null, Barisal, 8730, Bangladesh	Satisfactory
10	Bagerhat, Khulna, 9301, Bangladesh	Bagerhat, Khulna, Null, Bangladesh	Highly satisfactory

total number of missing words is M, and total number of incorrect words is W, then the accuracy of the recognition is computed as

$$\text{accuracy} = \left(1 - \frac{M + W}{T}\right) \times 100\%. \tag{1}$$

5.2. Evaluation Metrics. For the purpose of assessing how good the classifier is at predicting the class of the sample, we need to evaluate some performance measures. We derived a set of values from the confusion matrix, True Positive (TP) denotes the number of nonsuspicious samples that were correctly classified as nonsuspicious, and True Negative (TP) denotes the number of suspicious samples that were correctly classified as suspicious. False Positive (TP) denotes the number of suspicious samples incorrectly classified as nonsuspicious, and False Negative (TP) denotes the number of nonsuspicious samples incorrectly classified as suspicious. Based on these numbers, we evaluated several performance measures. Precision refers to the measure of exactness. It specifies what percentage of samples classified as nonsuspicious is actually nonsuspicious. Precision is evaluated by the following equation:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

Recall, also known as Sensitivity or True Positive Rate (TPR), is the measure of completeness. It specifies what percentage of nonsuspicious samples is classified as nonsuspicious. Precision can be calculated by the following formula:

$$\text{recall, sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

F_1 – score denotes the harmonic mean of Precision and Recall. The overall system performance can be depicted by the F_1 – score. It is calculated by the following formula:

$$F_1 - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{4}$$

5.3. Results. Our system classifies the input text into suspicious (C_S) and nonsuspicious (C_{NS}) classes based on the prediction probability. For the training of our LSTM model, we used 4000 samples and the remaining 1000 samples are used for validating the model as test data. Confusion matrix is a very useful tool for evaluating classification algorithms. As our problem is a binary classification, the confusion matrix is a 2×2 matrix. The model is evaluated by the 1000 samples and evaluated confusion matrix is shown in Table 3.

Our model was compared with other models based on other machine learning algorithms like Naive bayes, SVM, decision tree, k-nearest neighbor, and logistic regression. The comparison is summarized in Table 4. Naive Bayes is a simple probabilistic approach based on Bayes theorem [39]. By counting frequencies and combinations of values in the specified dataset, it calculates sets of probabilities and made the class prediction based on these probabilities. The working procedure of SVM is to find a maximum distant hyperplane between classes [40]. The support vectors create a hyperplane for binary classification that divides the cases into two nonoverlapping groups. We used Linear Kernel SVM for classification. Decision tree is a very popular algorithm in the field of text classification. It works by breaking down a set of data into smaller pieces. External nodes represent the class of decision, while internal nodes have the necessary features to render classification [41]. In our paper, we used CART (Classification and Regression Trees) algorithm for decision tree, which is very similar to

Table 2: Recognition accuracy of the framework for different languages.

Audio file	Actual speech	Converted text	Detected language	Number of missing words	Number of wrong words	Accuracy %
001.wav	আমি ভাল। আমি। তুমি কেমন আছো? তুমি মে ওখালন যালে?	আমি ভাল। আমি তুমি কেমন আছো তুমি মে এখালন যালে	Bengali	0	1	90
002.wav	Birds are flying in the sky.It seems so beautiful when they fly.	Birds are flying in the sky it seems so beautiful when they fly	English	0	0	100
003.wav	الْحَمْدُ ه لِلّ رَ ه بِ ا لعَالَمِينَ الرحمن الرحيم مالك يوم ال ه دين	الْحَمْدُ ه لِلّ رَ ه بِ ا لعَالَمِينَ الرحمن الرحيم مالك يوم ال ه دين	Arabic	0	0	100
004.wav	Fearlessness is like a muscle. I Know from my own life. ধন্যবাদে।	Hear is a music I know from my own life ধন্যবাদে।	Mixed	1	1	83.33
005.wav	بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ How are you? কতারিা মে ভাল। আলো?	بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ How are you কতারিা আলো	Mixed	2	p	84.61

Table 3: Confusion matrix.

Confusion matrix	Predicted nonsuspicious	Predicted suspicious
Actual nonsuspicious	True positive TP = 675	False negative FN = 31
Actual suspicious	False positive FP = 19	True negative TN = 275

Table 4: Comparison of performance.

Classification algorithm	Accuracy	Error	Precision	Recall	F_1-score
Naive Bayes	0.87	0.13	0.90	0.92	0.91
SVM	0.90	0.10	0.94	0.91	0.93
Decision tree	0.88	0.12	0.95	0.87	0.92
k-nearest neighbor	0.71	0.29	0.78	0.82	0.80
Logistic regression	0.92	0.08	0.96	0.91	0.93
Proposed LSTM based model	**0.94**	**0.06**	**0.96**	**0.95**	**0.95**

C4.5. We used Gini impurity measure as the function to measure the quality of split. KNN simply assigns the class of an unknown sample object by considering the majority of votes among the *k*-nearest neighbors. Logistic regression is a suitable choice for binary classification. It classifies a given sample into one of two classes [42]. All of these algorithms are trained and tested with the same dataset and using same train-test ratio (80 : 20).

From the comparison, we can see that our proposed LSTM based model performs much better in terms of accuracy. The main difference between our proposed LSTM based model and other machine learning models such as logistic regression, SVM, KNN, and Naive Bayes is the way of learning from data. For general machine learning methods like logistic regression, SVM, KNN, and Naive Bayes, we need to extract the features from the texts before applying the algorithm. For feature extraction from documents, we use TF-IDF (term frequency-inverse document frequency) vectorizer as it offers a way to determine the significance of the word on the basis of how much it appears in different documents [43]. However, these algorithms are highly dependent on the frequency of the words and fail to remember things in the past in an efficient manner. As text is sequential data, LSTM suits best in performing this task by remembering long-term dependencies within the sentence. Among all these algorithms, *k*-nearest neighbor performs poorly in terms of accuracy because KNN performs classification based on majority voting instead of learning from data.

6. Conclusion and Future Research

In this study, we proposed a framework for acquisition and detection of suspicious Bangla speeches. Bangla is one of the most spoken languages in the world[8], but, to the best of our knowledge, no work was done to detect suspicious Bangla speech. Our proposed system can classify Bangla speech into suspicious and nonsuspicious categories. As there is no such dataset available, we developed a dataset that contains 5000 samples. We developed an Android application for the acquisition of the speech. The application stores the speech and converts the speech into text. Our proposed framework for classifying suspicious speech was evaluated on the test data and a comparison with other machine learning algorithms like Naive Bayes, SVM, decision tree, *k*-nearest neighbor, and logistic regression was made. Among these, our proposed LSTM based model performs better in terms of accuracy.

There are some scopes for future research in our work. The recognition accuracy of multilingual (when a speech

contains multiple languages) speeches can be improved. Currently, our system can recognize speeches in Bangla, English, and Arabic languages but can classify speeches only in Bangla language; in the future, this classification can be enhanced for English and Arabic languages as well. Moreover, the dataset can be enriched by incorporating more speech samples from various sources. The classification categories can be increased to classify suspicious speeches into more specific classes instead of binary classification. Multiclass classification will make the detection much more precise.

With the rapid growth of access to Internet and online platforms, the misapplications of technologies have also increased rapidly. Ill-intentioned speeches can create unwanted situations in the society. An automatic tool that can detect suspicious speeches benefits governments and social network platforms to prevent unexpected situations.

Data Availability

The data cannot be made available on websites for public use due to some restrictions. However, the data can be collected for further research upon request to the first author via email: rsdrcse14@gmail.com.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded and supported by the project "Establishment of IT Business Incubator at CUET" via Reference no. 56.02.0000.023.16.032.18.93 (date: 06/11/2018).

References

[1] P. Suedfeld, S. Bluck, E. J. Ballard, and G. Baker-Brown, "Canadian federal elections: motive profiles and integrative complexity in political speeches and popular media," *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*, vol. 22, no. 1, pp. 26–36, 1990.

[2] H. Holzer, *Lincoln at Cooper Union: The Speech that Made Abraham Lincoln President*, Simon and Schuster, New York, NY, USA, 2004.

[3] M. Vail, "The "integrative" rhetoric of Martin Luther King Jr.'s "I have a dream" speech," *Rhetoric & Public Affairs*, vol. 9, no. 1, pp. 51–78, 2006.

[4] K. Gelber, "Terrorist-extremist speech and hate speech: understanding the similarities and differences," *Ethical Theory and Moral Practice*, vol. 22, no. 3, pp. 607–622, 2019.

[5] K. Gelber and L. McNamara, "Evidencing the harms of hate speech," *Social Identities*, vol. 22, no. 3, pp. 324–341, 2015.

[6] J. Waldron, *The Harm in Hate Speech*, Harvard University Press, Cambridge, MA, USA, 2012.

[7] E. Barendt, *Freedom of Speech*, Oxford University Press, Oxford, UK, 2nd edition, 2005.

[8] A. Buyse, "Words of violence: "fear speech," or how violent conflict escalation relates to the freedom of expression," *Human Rights Quarterly*, vol. 36, no. 4, pp. 779–797, 2014.

[9] A. K. Chen, "Free speech and the confluence of national security and internet exceptionalism," *Fordham Law Review*, vol. 86, no. 2, pp. 379–399, 2017.

[10] M. Islam and M. Islam, "Islam, politics and secularism in Bangladesh: contesting the dominant narratives," *Social Sciences*, vol. 7, no. 3, p. 37, 2018.

[11] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 85–94, Prague, Czech Republic, July 2017.

[12] N. Gandhewar and R. Sheikh, "Google Android: an emerging software platform for mobile devices," *International Journal on Computer Science and Engineering*, vol. 1, no. 1, pp. 12–17, 2010.

[13] R. E. Rice and J. E. Katz, "Comparing internet and mobile phone usage: digital divides of usage, adoption, and dropouts," *Telecommunications Policy*, vol. 27, no. 8-9, pp. 597–623, 2003.

[14] P. Kaur and S. Sharma, "Google Android a mobile platform: a review," in *Proceedings of the 2014 Recent Advances in Engineering and Computational Sciences (RAECS)*, pp. 1–5, Chandigarh, India, March 2014.

[15] G. S. Chavan, S. Manjare, P. Hegde, and A. Sankhe, "A survey of various machine learning techniques for text classification," *International Journal of Engineering Trends and Technology*, vol. 15, no. 6, pp. 288–292, 2014.

[16] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.

[17] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web—WWW'16*, pp. 145–153, Montreal, Canada, May 2016.

[18] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78, 2019.

[19] O. Oriola and E. Kotze, "Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020.

[20] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in Tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion—WWW'17 Companion*, pp. 759-760, Perth, Australia, April 2017.

[21] B. Gambäck and U. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proceedings of the First Workshop on Abusive Language Online*, pp. 85–90, Vancouver, Canada, August 2017.

[22] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a Convolution-GRU based deep neural network," in *Proceedings od the The Semantic Web. ESWC 2018*, pp. 745–760, Heraklion, Greece, June 2018.

[23] J. Risch and R. Krestel, "Toxic comment detection in online discussions," in *Algorithms for Intelligent Systems*, pp. 85–109, Springer, Singapore, 2020.

[24] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, 2020.

[25] S. M. A. Taher, K. A. Akhter, and K. M. A. Hasan, "N-gram based sentiment mining for Bangla text using Support Vector Machine," in *Proceedings of the 2018 International Conference*

on Bangla Speech and Language Processing (ICBSLP), pp. 1–5, Sylhet, Bangladesh, September 2018.

[26] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," in *Proceedings of the 16th International Conference on Computer and Information Technology*, pp. 366–371, Khulna, Bangladesh, March 2014.

[27] A. Dhar, N. Dash, and K. Roy, "Classification of Bangla text documents based on inverse class frequency," in *Proceedings of the 2018 3rd International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pp. 1–6, Bhimtal, India, February 2018.

[28] O. Sharif and M. Hoque, "Automatic detection of suspicious Bangla text using logistic regression," in *Proceedings of the International Conference on Intelligent Computing & Optimization*, pp. 581–590, Koh Samui, Thailand, December 2019.

[29] A. M. Ishmam and S. Sharmin, "Hateful speech detection in public Facebook pages for the Bengali language," in *Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 555–560, Boca Raton, FL, USA, December 2019.

[30] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra, "A deep learning approach to detect abusive Bengali text," in *Proceedings of the 2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pp. 1–5, Sarawak, Malaysia, June 2019.

[31] T. Islam, S. Latif, and N. Ahmed, "Using social networks to detect malicious Bangla text content," in *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–4, Dhaka, Bangladesh, May 2019.

[32] P. Chakraborty and M. H. Seddiqui, "Threat and abusive language detection on social media in Bengali language," in *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–6, Dhaka, Bangladesh, May 2019.

[33] M. Dwarampudi and N. V. S. Reddy, *Effects of Padding on LSTMs and CNNs*, https://arxiv.org/abs/1903.07288, 2019.

[34] G. Weiss, Y. Goldberg, and E. Yahav, "On the practical computational power of finite precision RNNs for language recognition," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 740–745, Melbourne, Australia, July 2018.

[35] D. Britz, A. Goldie, M. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1442–1451, Copenhagen, Denmark, September 2017.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations*, San Diego, CA, USA, May 2015.

[38] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[39] S. Wang and C. D. Manning, "Baselines and bigrams: simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of The Association for Computational Linguistics: Short Papers-Volume 2*, pp. 90–94, Jeju Island, South Korea, July 2012.

[40] M. Karan and J. Šnajder, "Cross-domain detection of abusive language online," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 132–137, Brussels, Belgium, October-November 2018.

[41] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[42] E. F. Unsvag and B. Gambäck, "The efects of user features on twitter hate speech detection," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 75–85, Brussels, Belgium, October-November 2018.

[43] J. Salminen, H. Almerekhi, M. Milenković, S. Jung, J. An et al., "Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2018)*, pp. 330–339, Palo Alto, CA, USA, June 2018.

Hindawi
Complexity
Volume 2020, Article ID 9757032, 14 pages
https://doi.org/10.1155/2020/9757032

Research Article

A Deep Paraphrase Identification Model Interacting Semantics with Syntax

Leilei Kong, Zhongyuan Han, Yong Han, and Haoliang Qi

School of Electronic Information Engineering, Foshan University, Foshan 528225, China

Correspondence should be addressed to Zhongyuan Han; hanzhongyuan@gmail.com

Received 13 July 2020; Revised 25 September 2020; Accepted 4 October 2020; Published 30 October 2020

Academic Editor: Abd E. I.-Baset Hassanien

Paraphrase identification is central to many natural language applications. Based on the insight that a successful paraphrase identification model needs to adequately capture the semantics of the language objects as well as their interactions, we present a deep paraphrase identification model interacting semantics with syntax (DPIM-ISS) for paraphrase identification. DPIM-ISS introduces the linguistic features manifested in syntactic features to produce more explicit structures and encodes the semantic representation of sentence on different syntactic structures by means of interacting semantics with syntax. Then, DPIM-ISS learns the paraphrase pattern from this representation interacting the semantics with syntax by exploiting a convolutional neural network with convolution-pooling structure. Experiments are conducted on the corpus of Microsoft Research Paraphrase (MSRP), PAN 2010 corpus, and PAN 2012 corpus for paraphrase plagiarism detection. The experimental results demonstrate that DPIM-ISS outperforms the classical word-matching approaches, the syntax-similarity approaches, the convolution neural network-based models, and some deep paraphrase identification models.

1. Introduction

The goal of paraphrase identification is to determine whether two texts have the same meaning [1]. It focuses on how best to model the semantics of sentences [2]. Paraphrase identification is one of the most basic problems in lots of applications of natural language processing, such as machine translation [3], question and answering [4], plagiarism detection [5,6], and document retrieval [7].

Although paraphrase identification is commonly defined in semantic terms [2], the early methods to paraphrase identification were usually based on the word (or word n-gram) matching or the vector similarity in the word space, without considering the semantics of words or sentences. The bag-of-words model [8], the n-gram model [9], the TFIDF [10] (term frequency and inverse document frequency) model, and so on were commonly applied to represent the text, and then some text similarity computing methods (such as edit distance, longest common substring, Jaccard coefficient, and cosine distance) were exploited to measure the degree of paraphrase between the two texts. However, paraphrase is usually done by word replacement with synonyms/antonyms, syntactic modification, sentence reduction, combination, reorganization, word shuffling, concept generalization, and specificity to change the appearance of the original text while retaining the semantics of the source sentence [11], which makes the above methods difficult to further improve the performances using only word matching or vector similarity in the word space.

The syntactic feature-based methods, another way without considering the semantics, have also been used in paraphrase identification [11–13], especially in cross-language paraphrase identification [14]. These studies assume that similar texts have similar syntactic structures [12, 15]. That is, if two sentences describe the same thing, they are likely to have similar syntactic structures [16]. However, simply relying on the similarity of syntactic structures without regard to semantics cannot solve the problem of "the same semantics but different syntactic structures" [17].

In recent years, the models of paraphrase identification tend to transfer from the traditional model to deep model [18]. A variety of deep models have been introduced into the

research field of paraphrase identification [19–24]. These models utilized the distributed representation of text and focused on identifying the paraphrase through learning the matching structures and the matching degrees.

Except for the widely accepted distributed semantic representation in the deep paraphrase identification models, researchers also paid attention to the role of syntax in representing the text and computing the semantic similarity, and proposed some deep paraphrase identification models integrating the syntax [16, 25]. These studies determined the validity of syntactic features in deep paraphrase identification.

Goldberg presented that the linguistic features providing the more explicit general concepts can be very valuable [26]. Hu et al. proposed that a successful sentence-matching algorithm needs to capture not only the internal structures of sentences but also the rich patterns in their interactions [21]. We deem that the linguistic features manifested in syntactic features can produce more explicit structures for the representation of sentences and modeling the semantics on these syntactic features by means of the interaction of semantics with syntax can better represent the sentences and help to identify paraphrase.

Based on this, we propose a novel deep paraphrase identification model interacting semantics with syntax, denoted as DPIM-ISS. DPIM-ISS represents the sentences as the semantic vector on syntactic features and characterizes the syntactic role for the semantics of word or phrases by interacting semantic and syntactic information. Exploiting this representation, DPIM-ISS models the semantic representation on syntactic features explicitly and permits the model to learn the paraphrase pattern from the semantic on different linguistic features.

DPIM-ISS is evaluated on three datasets: MSRP (Microsoft Research Paraphrase) [27], PAN 2010 [6], and PAN 2012 [28]. The experimental results show that the proposed model outperforms the traditional word-matching approaches, the syntax-similarity approaches, the distributed-representations-of-sentences-based models, the CNN-based models, and a couple of deep models for paraphrase identification.

The contributions of this paper can be summarized as follows:

(i) The idea of modeling the semantic representation of sentence on different syntactic structures by means of interacting semantics with syntax

(ii) A new application of deep architecture, namely DPIM-ISS, to exploit the sentence representation interacting semantics with syntax for paraphrase identification

(iii) Experiments on three datasets (i.e., MSRP, PAN 2010, and PAN 2012) to show the benefits of our model

The following sections are organized as follows: Section 2 analyzes the issues of paraphrase identification. Section 3 introduces the details of DPIM-ISS. The experimental results are reported in Section 4. Section 5 discusses the related work. Section 6 concludes our work.

2. Analysis of Paraphrase Identification

Taking the data of MSRP and PAN (the detailed statistics of the two datasets can be found in Section 4.1) as examples, we investigate the semantic similarity of the sentences from the aspects of lexical similarity and syntactic similarity to denote the paraphrase.

2.1. Paraphrase Sentences with High Lexical Similarity. From the perspective of word matching, the sentences are more than likely being paraphrased if they use the same or similar words. We randomly selected 1000 pairs of paraphrase sentences and 1000 pairs of nonparaphrase sentences from the MSRP dataset and compared their lexical similarity using Jaccard coefficient, as Figure 1 shows.

Figure 1 reveals that when Jaccard coefficient is higher than 0.6, most of the sentence pairs are paraphrase sentences, while when Jaccard coefficient is lower than 0.25, most of the sentence pairs are nonparaphrase sentences.

Analyzing the examples of paraphrase sentences, we find that if the paraphrase sentences rewrite the source sentences by simple duplication, the syntactic structures of the two sentences are the same or similar, while if the paraphrase sentences rewrite the source sentences by text manipulation such as adjusting word orders or modifying the syntactic structures, the syntactic structures of the two sentences will therefore be different, but the words are still the same or similar. It shows that the word matching is still valuable in the paraphrase identification task. When Jaccard coefficients are between 0.25 and 0.6, it is difficult to distinguish paraphrase or nonparaphrase.

2.2. Paraphrase Sentences with the Same (Similar) Syntactic Structures but Different Words. From the view of the syntactic structure, some paraphrase sentences have the same or similar syntactic structures but different words. Figure 2 gives a pair of paraphrase sentence from PAN 2012 with low lexical similarity but high syntactic similarity.

Figure 2 exemplifies a lexical paraphrase, where underlined words are replaced with synonyms, and short phrases or words are inserted to change the appearance of the text. Although much of the text is changed, paraphrasing retains the semantics of the source. It is a common type of case in paraphrase identification. The higher the degree of paraphrase, the more difficult to identify paraphrase only by word matching.

If the word matching is not considered and only the syntactic features are exploited, the pairs of such paraphrase sentences are more similar on syntactic structures. Figure 3 compares the Jaccard coefficients of syntactic features computed from 1000 pairs of paraphrase and nonparaphrase sentences randomly selected from the training dataset of MSRP. The X-axis records the Jaccard coefficients, and the y-axis is the number of the samples.

The statistical information in Figure 3 shows that the number of paraphrase sentence pairs is significantly higher than that of nonparaphrase sentence pairs as the similarity of the syntactic feature sequence increases. For example, when

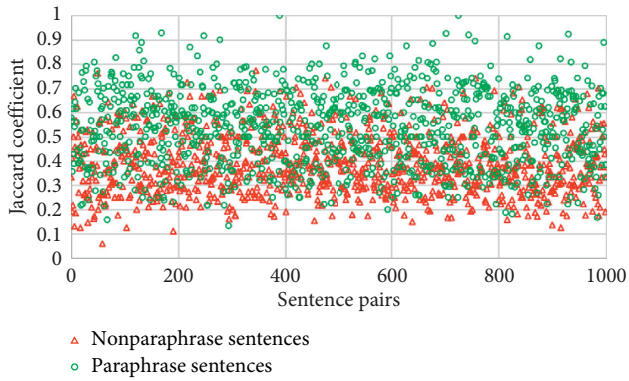


Figure 1: Comparisons of lexical similarity between paraphrase pairs and nonparaphrase pairs.

Source sentence: In the early 1800's doc bowles built the first hotel, a three story frame building. The community thrived and there was an influx of tourist traffic coming to drink and soak in the mineral waters. In the 1850's French lick was a key station in the "underground railway". The French lick springs resort and spa was built in the late 1800's.
Paraphrased sentence: In an French[insertion] 1800's tourist[substitution] bowles built a[substitution] first drink[substitution], the[substitution] three hotel railway mineral[substitution]. The community thrived and there was the[substitution] influx in[synonym] doc[substitution] traffic coming in story[substitution] and soak in the building[substitution] waters. To[synonym] the 1850's French lick was the early[substitution] station in the "underground frame[synonym] ". The key[substitution] lick springs resort and spa was built of a late[substitution] 1800's.

Figure 2: The samples of paraphrase sentence pair: different words, the similar syntactic structure, and the same semantics.

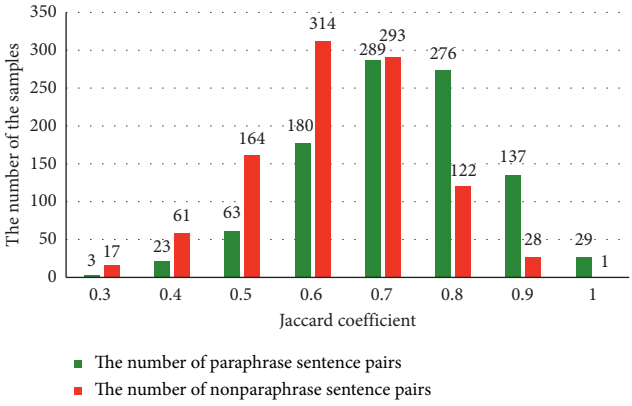


Figure 3: Comparison of syntactic similarity between paraphrase pairs and nonparaphrase pairs using Jaccard coefficient.

the Jaccard coefficients of the sentence pairs are between 0.8 and 0.9, there are 137 pairs of paraphrase sentences and only 28 pairs of nonparaphrase sentences. Therefore, the similarity of the syntactic structure is useful to the task of paraphrase identification.

2.3. Nonparaphrase Sentences with Similar Words and Similar Syntax Structures. Figure 4 describes an example of nonparaphrase sentences with similar words (black part) and similar syntax structures (see the dependency parsing tree corresponding to two sentences). In this example, S_1 and S_2 share a large number of the same words. Without respect to the semantics, the two sentences will be recognized as paraphrase due to the high levels of word matching. Similarly, S_1 and S_2 can be identified as paraphrase since they have basically the same syntactic structures. However, if we compare the semantics of the words defined on the dependency tree, we can find that the semantics of verb appeared and surrendered are completely different, which leads to the semantic difference between the two sentences.

2.4. Different Words and Different Syntax, but the Same Semantics. Figure 5 shows an example in MSRP corpus with different words and different syntactic structure, but the same semantics.

In Figure 5, there are few identical words between two paraphrase sentences and the syntactic structures are much more varied. However, if we map the semantics of words to the substructures expressed by the dependency tree of sentences and compare the semantics of words in the syntactic substructures, such as refused and denied on VBD, the semantic similarity of the two sentences can be found.

A sentence written in the natural language is not the simple collection of words, but the text with the syntactic structure under the grammar restriction.

There exist the corelationships between semantics and syntax: when we need to convey and express the message in a proper way, the semantics and syntax of the sentence will work together, which encourages us to interact syntax and semantics in paraphrase identification to boost the performance.

3. Deep Paraphrase Identification Model Interacting Semantics with Syntax

The architecture of the deep paraphrase identification model interacting semantics with syntax (DPIM-ISS) contains two components: the sentence representation interacting semantics with syntax and the extraction of the matching pattern based on convolutional neural network. In this section, we introduce DPIM-ISS in detail.

3.1. Overview of DPIM-ISS. Paraphrase identification is usually formalized as a binary classification task [29]: given two sentences (s_k, s_p), the paraphrase identification model M determines whether they roughly have the same meaning. We propose DPIM-ISS to learn M, as shown in Figure 6.

In the architecture of DPIM-ISS illustrated in Figure 6, the model contains the two main parts: (1) the sentence representation interacting semantics with syntax, and (2) the extraction of the paraphrase matching pattern based on convolutional neural network. In what follows, we describe these components in detail.

3.2. The Sentence Representation Interacting Semantics with Syntax. In recent years, the tensor has attracted much attention due to its ability to model the interaction between objects. For example, Socher et al. proposed a neural tensor network to model the interaction of two entities [30] and Qiu et al. modeled the interaction between the questions and answers using tensor in the task of community question

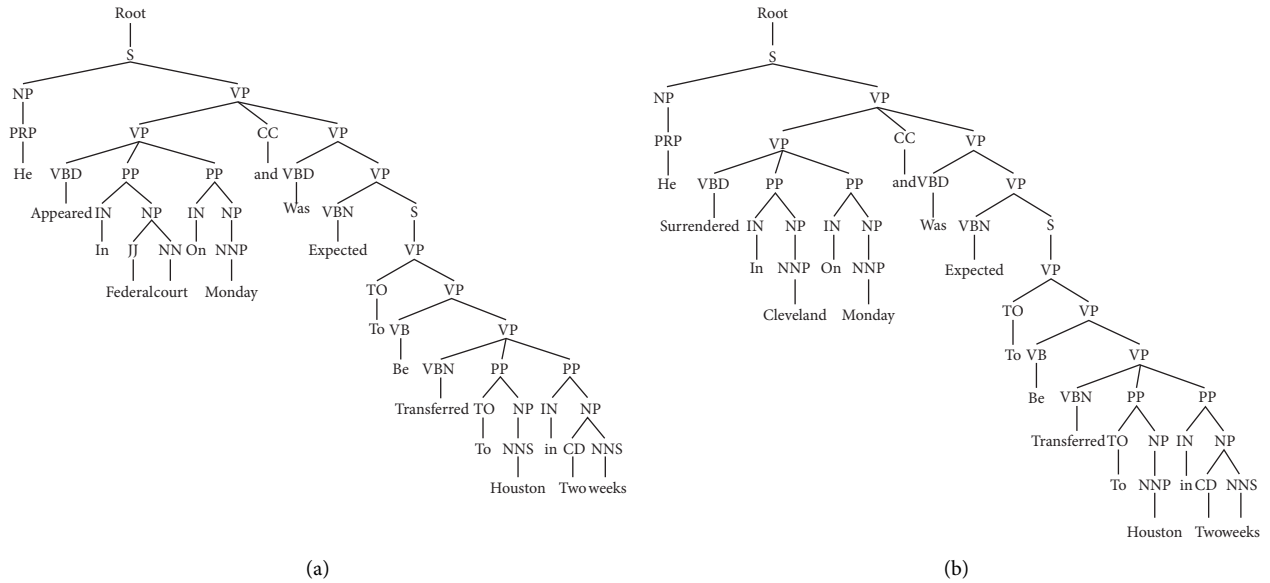


Figure 4: Nonparaphrase sentence pair-similar words with similar syntax structure. (a) S_1: He appeared in federal court on Monday and was expected to be transferred to Houston in two weeks. (b) S_2: He surrendered in Cleveland on Monday and was expected to be transferred to Houston in two weeks.

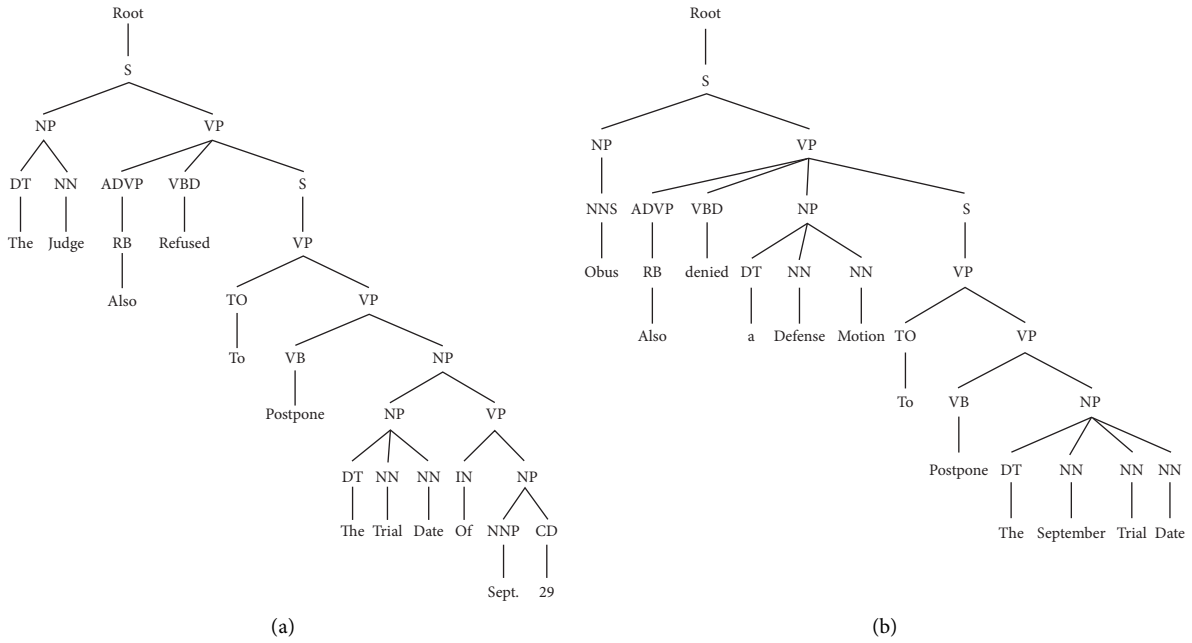


Figure 5: Paraphrase sentence pair-different words with different syntax structure. (a) S_1: The judge also refused to postpone the trial date of Sept. 29. (b) S_2: Obus also denied a defense motion to postpone the September trial date.

answering [31]. In the study of Yu et al., the idea of tensor was exploited to model the interaction between the semantic information and the structural information [32]. The motivations of these methods are all to use tensor as the tool to capture the interaction between different features. Inspired by these studies, DPIM-ISS uses tensor to interact the semantics and syntactic structures to model the sentence representation. Figure 7 gives a detailed example.

Given a sentence $s_k = \{w_1^{(k)}, \ldots, w_i^{(k)}, \ldots, w_n^{(k)}\}$, where $w_i^{(k)}$ is the i-th word of s_k, let $e_{w_i}^{(k)}$ denote the semantic feature vector of $w_i^{(k)}$ represented as word embedding and $g_{w_i}^{(k)}$ be the syntax feature vector of $w_i^{(k)}$ that provides the syntax role of $w_i^{(k)}$. DPIM-ISS uses the tensor product $\otimes$ of $e_{w_i}^{(k)}$ and $g_{w_i}^{(k)}$ to project the structure of interacting semantics with syntax for the word $w_i^{(k)}$, represented by the notation $x_{w_i}^{(k)}$:

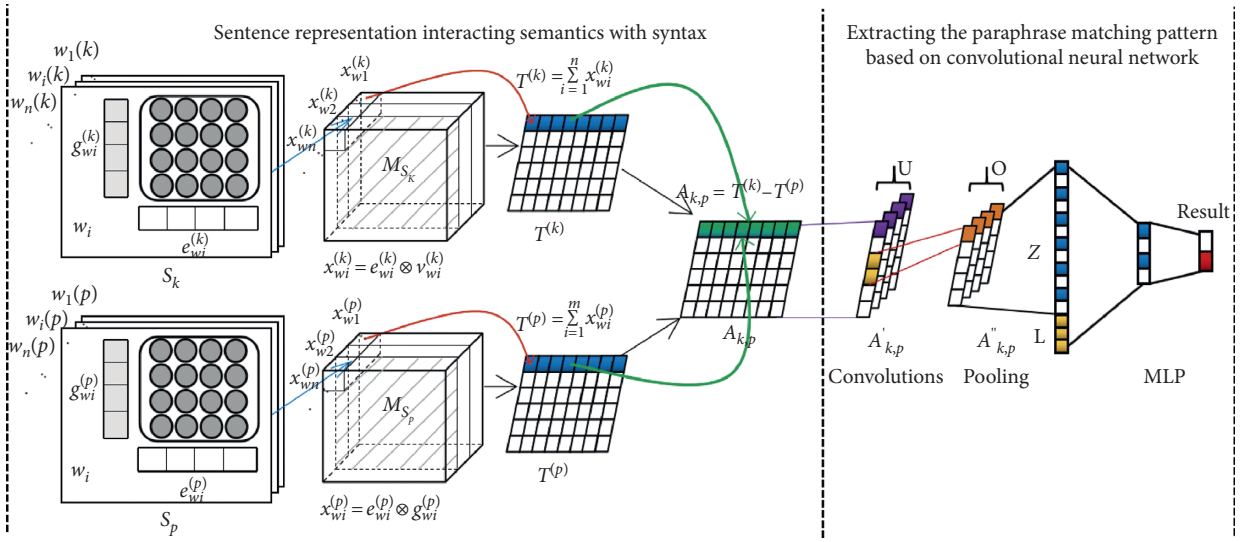


Figure 6: The architecture of the deep paraphrase identification model interacting semantics with syntax (DPIM-ISS).

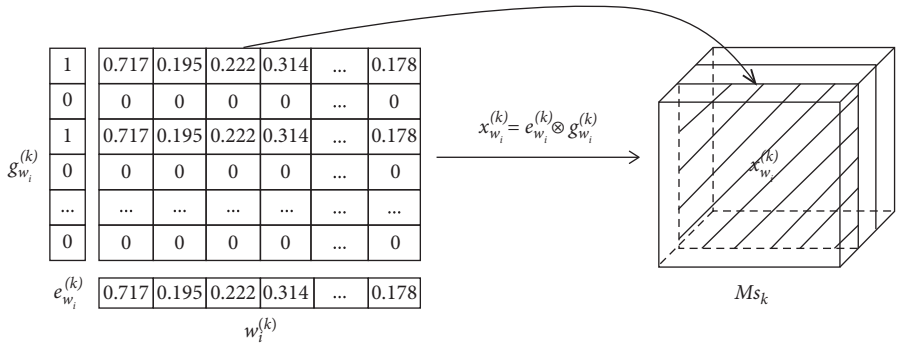


Figure 7: Interacting semantics with syntax.

$$x_{w_i}^{(k)} = e_{w_i}^{(k)} \otimes g_{w_i}^{(k)}. \tag{1}$$

Let $g_{w_i}^{(k)} = \{g_1, g_2, \ldots, g_m, \ldots, g_M\}$ denote an example of predefined syntax feature vector template of size M. Each $g_{(m)}$ represents a fixed syntax feature such as the subject (the syntactic component), the noun (the speech), and so on. Given a word $w_i^{(k)}$, $g_{w_i}^{(k)}$ is a binary vector to represent the categorical variables. The categorical values are mapped to all zero values except those syntactic features that $w_i^{(k)}$ has. We use the syntactic parsing to obtain the syntax feature $g_{w_i}^{(k)}$ for the k-th word $w_i^{(k)}$ in sentence s. For example, in $g_{w_i}^{(k)}$, $g_{(m)} = 1$ means that $w_i^{(k)}$ has the m-th syntactic feature; on the contrary, $g_{(m)} = 0$ indicates that $w_i^{(k)}$ does not act as the role of m-th syntactic feature.

Using the semantic feature vector $e_{w_i}^{(k)}$ and the syntactic feature vector $g_{w_i}^{(k)}$, we then generate the word embedding representation $x_{w_i}^{(k)}$ interacting the semantics with syntax using equation (1). Each $x_{w_i}^{(k)}$ is a two-dimensional matrix, shown on the left of Figure 7. Furthermore, if each word in the sentence s_k is represented using the word embedding interacting the semantics with syntax, then we can get a three-dimensional matrix M_{s_k} to represent the sentence s_k, shown on the right of Figure 7.

M_{s_k} consists of three dimensions: each word w in sentence s, the semantic feature vector e, and the syntactic feature vector g. DPIM-ISS captures the interactions between semantic features and syntactic features using tensor product, depicts the semantics of words on syntactic roles and decomposes the sentence into the syntactic subsections with semantics.

In order to obtain the expression of a sentence, we sum the word embedding interacting semantics with syntax to map M_{s_k} into a two-dimensional space of semantic and syntactic dimensions, shown in equation (2). Then, we obtain the representation $T^{(k)}$ of the sentence s_k, called the sentence representation interacting semantics with syntax in this paper:

$$T^{(k)} = \sum_{w_i \subset S_k} x_{w_i}^{(k)}. \tag{2}$$

Furthermore, given two sentences s_k and s_p, we represent the interaction between them as a vector $A_{k,p}$ as follows:

$$A_{k,p} = T^{(k)} - T^{(p)}. \tag{3}$$

Then, the feature vector $A_{k,p}$ is further fed to a convolutional neural network to extract the paraphrase matching pattern.

3.3. Extracting the Paraphrase Matching Pattern Based on Convolutional Neural Network. The convolutional neural networks have been applied to learn effective feature representations in some language tasks in recent years. In DPIM-ISS, we use the convolutional neural networks to extract the features of paraphrase matching. Then, the extracted features will be fed into a multilayer perceptron classifier to identify the paraphrase.

3.3.1. Convolutional Layer. We use wide one-dimensional convolution [33], which was proposed by Kalchbrenner et al., to define the convolution kernel to extract the features from $A_{k,p}$ for paraphrase identification. In DPIM-ISS, $A_{k,p}$ is the interacting representation between the two sentences, and it is an $m \times n$ matrix, where m is the number of syntactic features and n is the dimension of semantic features.

The convolution layer exploits the U convolution kernels of size $1 \times n$ and a convolution kernel contains two parameters: W and b, where $W = [w_1, \ldots, w_n]$ is the feature weight vector of the convolution kernel and b is the bias of the convolution kernel. A convolution kernel performs the convolutional operation on the interaction matrix $A_{k,\mathrm{p}}$ to get an $m \times 1$ vector V_u, which represents the expression of a semantic feature on a syntactic feature. V_u is defined as follows:

$$V_u = \mathrm{Cov}\left(W, A_{k,p}\right) + b, \tag{4}$$

where $\mathrm{Cov}(W, A_{k,p})$ denotes a convolution operation on $A_{k,\mathrm{p}}$ using parameter W:

$$\mathrm{Cov}\left(W, A_{k,p}\right) = \begin{bmatrix} \sum_{i=1}^{n} w_i * A_{1,i} \\ \cdots \\ \sum_{i=1}^{n} w_i * A_{m,i} \end{bmatrix}. \tag{5}$$

The convolution operation explores U convolution kernels to produce a matrix $A'_{k,p} \in R^{U \times m \times 1}$, which is composed by $m \times 1$ vector $V^{(u)} = [v_1^{(u)}, v_2^{(u)}, \ldots, v_m^{(u)}]^T$, where m is the number of syntactic features and U is the number of convolution kernels:

$$A'_{k,p} = \left[V^{(1)}, V^{(2)}, \ldots, V^{(u)}, \ldots, V^{(U)}\right]. \tag{6}$$

3.3.2. Max Pooling. The outputs from the convolutional layer are then passed to the pooling layer to extract the k top values from each dimension of $A'_{k,p}$ for reducing the number of the features. On each column of $A_{k,p}{}'$, we set the size of nonoverlapping pooling window to w. The k features with the highest value are extracted from the window, and the matrix $A''_{k,p} \in R^{(U \times m/w \times 1)}$ made up of k m/w vectors is generated as follows:

$$A''_{k,p} = \left[V''^{(1)}, V''^{(2)}, \ldots, V''^{(u)}, \ldots, V''^{(U)}\right], \tag{7}$$

where each $V''^{(u)}$ is defined as follows:

$$V''^{(u)} = \begin{bmatrix} \max\left(v_1^{(u)}, \ldots, v_w^{(u)}\right) \\ \max\left(v_{w+1}^{(u)}, \ldots, v_{w+w}^{(u)}\right) \\ \cdots \\ \max\left(v_{\lfloor m/w \rfloor * w+1}^{(u)}, \ldots, v_m^{(u)}\right) \end{bmatrix}. \tag{8}$$

Then, the resulting features of $A''_{k,p}$ operated by max pooling are combined to form a $k \times m/\mathrm{w}$ dimensional vector Z.

3.3.3. Further Enhancements. Madnani et al. proved that the machine translation (MT) metrics significantly boosted the performance of paraphrase identification [6]. For each pair of sentences, we construct a vector L to indicate the lexical similarity using the METEOR automatic MT evaluation metric, including precision, recall, F1, Fmean, penalty, and METEOR score [34]. We refer to such vector as the lexical features and incorporate it into the proposed DPIM-ISS by appending it to the vector Z. We conducted several experiments both with and without these features, which are discussed below.

3.3.4. Identifying Paraphrase. We pass Z with L to a two-layer perceptron, shown in equation (9):

$$\left(p_0, p_1\right)^T = \delta_2\left(W_2 \delta_1\left(W_1 Z + b_1\right) + b_2\right), \tag{9}$$

where p_0 and p_1 indicate the identification results, W_i and b_i are the weight matrix and the bias of the i-th layer of the perceptron, respectively, and δ_1 is the ReLU activation function [35], defined as follows:

$$f(x) = \max(0, x), \tag{10}$$

and δ_2 is the SoftMax function to output the value of p_k:

$$p_k = \frac{e^{ak}}{e^{a_0} + e^{a_1}}, \quad k = 0, 1, \tag{11}$$

where a_k is the output value after ReLU activation function in the last layer.

3.3.5. Training the Model. During the training phase, parameters of DPIM-ISS are updated with respect to a cross-entropy loss between the predicted results and the ground truth, and the regulation technology is adopted to avoid the overfitting problem. The loss function is defined as follows:

$$C(W, b) = -\frac{1}{N} \sum_{M_i} \left[y^{(i)} \log\left(p_1^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(p_0^{(i)}\right)\right] + \frac{\lambda}{2N} \sum_{W \in \{W_1, W_2\}} W^2, \tag{12}$$

where $y^{(i)}$ is the label of i-th training example, λ is the regularization coefficient, and W_1 and W_2 are the parameters of the two-layer perceptron.

To train the model, we use the backpropagation algorithm [36] with the Adam update rule [37]. The updating forms of parameters are as follows:

$$W_t = W_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \varepsilon}, \tag{13}$$

where t is the current timestep, W_t is the weights of t-th timestep, W_{t-1} is the weights of W in the last round of training, η is the learning rate, ε is a parameter, and $\widehat{m}_t$ and $\widehat{v}_t$ are the bias-corrected estimates to control the direction of the gradient. We set $\eta = 1e{-}4$ and $\varepsilon = 1e{-}08$.

The whole sentence representation interacting semantics with syntax and the training process are detailed in Appendix A.

4. Experiments

4.1. Datasets. We conduct our experiments on three datasets: the Microsoft Research Paraphrase (MSRP) [27], the PAN 2010 [6], and the PAN 2012 [28]. MSRP is a classical dataset for paraphrase identification developed by Microsoft, and the latter two datasets are constructed using the datasets of 2010 and 2012 Uncovering Plagiarism, Authorship and Social Software Misuse shared task.

4.1.1. MSRP. The MSRP corpus is a well-known corpus for paraphrase identification. MSRP was created by mining the news articles on the web and then extracting the paraphrases sentences from 9,516,684 sentences in 32,408 news clusters by using a semiautomatic method. It contains 5,801 sentential pairs, which is split into 4,076 (2,753 paraphrase, 1,323 not) training and 1,725 (1,147 paraphrase, 578 not) test pairs.

4.1.2. PAN 2010. Madnani and Tetreault used the human-created plagiarism instances in the test collection from the PAN 2010 plagiarism detection competition to create the PAN 2010 paraphrase sentence corpus. They utilized the bag-of-words overlap and length ratios to generate the pairs of paraphrase sentences and selected the sentence pairs that had at least 4 words in common from the same document as the pairs of nonparaphrase sentence. Then, they sampled randomly from both the positive and negative instances to create a training set of 10,000 sentence pairs and a test set of 3,000 sentence pairs.

4.1.3. PAN 2012. We constructed the PAN 2012 paraphrase sentence pair dataset using the training and test data of PAN 2012 paraphrase plagiarism detection corpus. Let d_{plg} and d_{src} denote the plagiarized document and its source document, and (s, r) is a pair of plagiarism text annotated by PAN ($s \in d_{\text{plg}}, r \in d_{\text{src}}$). Let $s_i \in s$ be the sentence of s and $r_j \in d_{src}$ denote the sentence of d_{src}, and $T = \{y, (s_i, r_j)\}$ represent the training dataset. y and r_j are defined as follows:

$$\begin{cases} y = 1, & r_j = \arg\max_{r_j \in r} \left(\cos\left(s_i, r_j\right)\right), \\ y = 0, & r_j = \arg\max_{r_j \in d_{\text{src}}, r_j \notin r} \left(\cos\left(s_i, r_j\right)\right), \end{cases} \tag{14}$$

where $\cos(s_i, r_j)$ is the cosine similarity of s_i and r_j. Using the proposed method, we obtained 341,426, and 50,114 pairs of paraphrase sentences from the artificial-high-obfuscation subcorpus of PAN 2012 training and test corpus. Then 15,932 training pairs and 7,966 test pairs that the length ratios were more than 50% were randomly selected to generate our PAN 2012 paraphrase sentence pairs dataset.

The statistics of the three datasets are described in Table 1.

4.2. Experimental Setting

4.2.1. Baselines. We evaluate the effectiveness of our model with several baseline methods, including the traditional word-matching approaches, the syntax-similarity approaches, the distributed-representations-of-sentences-based models, and the CNN-based models. At the same time, we also select multiple deep paraphrase identification models as baselines. We give a detailed description of these baselines as follows:

(1) Word-Matching Approaches. We select four typical word-matching approaches as baselines.

Jaccard. The Jaccard method calculates the Jaccard coefficient of the two sentences first and selects those pairs whose Jaccard coefficients are greater than a threshold t as the paraphrase sentence pairs. In experimenting, we set t from 0.0 to 1.0 and let the incremental step length be 0.01. We selected the parameter t on the training corpus in terms of optimizing accuracy. Then, the corresponding t was applied on the test corpus. On the MSRP dataset, $t = 0.34$. On PAN 2010, $t = 0.24$. On PAN 2012, $t = 0.27$.

Cosine. Similar to the Jaccard method, the cosine method calculates the similarity of the two sentences using the cosine distance. Similar to the above Jaccard method, we set a threshold t to decide the paraphrase sentence pairs. On the MSRP dataset, $t = 0.28$. On PAN 2010, $t = 0.34$. On PAN 2012, $t = 0.20$.

METEOR. We applied the six METEOR evaluation metrics as the features to learn a classifier using the logic regression model (in DPIM-ISS, these lexical features are integrated into the extracted features that interact semantics with syntax). All parameters are obtained based on the training data to optimize F1.

(2) Syntax-Similarity-Based Approaches (Syntax-sim). For syntactic similarity, we referred to the method proposed in [11], denoted as Syntax-sim (Syntax-similarity). In Syntax-sim, we considered the text as the string of syntactical sequences derived from Stanford POS tagging[1] instead of using actual words and utilized the Jaccard coefficient to compute the similarity of syntactical sequences for further decision.

Table 1: The statistics of the datasets.

Datasets			Training data	Test data
MSRP	Number of sentence pairs		4076	1725
	Length of sentence pairs	Short ≤ 20 words	2.40%	2.78%
		Medium 20–50 words	86.83%	86.03%
		Long >50 words	10.77%	11.19%
		Max length	60	63
		Min length	14	12
	Jaccard coefficient	<3%	0.02%	0.00%
		3%–10%	0.02%	0.12%
		10%–30%	13.10%	13.86%
		30%–50%	42.93%	43.65%
		50%–80%	41.78%	40.12%
		>80%	2.13%	2.26%
PAN 2010	Number of sentence pairs		10000	3000
	Length of sentence pairs	Short ≤ 50 words	35.76%	35.80%
		Medium 50–200words	63.79%	63.93%
		Long >200 words	0.45%	0.27%
		Max length	477	272
		Min length	3	5
	Jaccard coefficient	<3%	0.24%	0.20%
		3%–10%	3.22%	2.43%
		10%–30%	57.71%	57.90%
		30%–50%	18.27%	18.13%
		50%–80%	18.52%	19.47%
		>80%	2.04%	1.87%
PAN 2012	Number of sentence pairs		15932	7966
	Max length	Short ≤ 50 words	51.82%	51.42%
		Medium 50–200words	46.94%	47.39%
		Long >200 words	1.24%	1.19%
		Max length	1833	1658
		Min length	22	22
	Min length	<3%	0.01%	0.01%
		3%–10%	3.25%	3.09%
		10%–30%	75.71%	75.77%
		30%–50%	18.25%	17.95%
		50%–80%	2.74%	3.09%
		>80%	0.05%	0.09%

(3) Distributed-Representations-of-Sentences-Based Model (Paragraph Vector). In our DPIM-ISS, we focus on the distributed representation of sentences. Thus, we select a distributed-representations-of-sentences-based model, the paragraph vector, proposed in [38] as the baseline for comparison. Paragraph vector used an unsupervised algorithm to learn the sentence representations. We utilized the tools of gensim[2] to learn the sentence vector and applied the cosine distance to compute the similarity of the two sentences. The parameter settings are as follows: the size of context window is 5, the lowest word frequency is 5, the learning rate is 0.025, and the dimension of sentence vector is 300.

(4) CNN-Based Models. ARC-I DPIM-ISS exploits the convolutional neural network to extract the paraphrase patterns of the interacting sentence representation. We also select a CNN-based paraphrase identification model, the ARC-I [21], as the baseline. In the experiment, we reimplemented ARC-I due to no publicly available codes, using the network structure and parameter setting as described in the original paper. The word embedding used for ARC-I was as the same as DPIM-ISS (will be described in 5.2.3). All parameters were obtained based on training data to optimize F1.

(5) Other Deep Paraphrase Identification Models. We also compared the performance of DPIM-ISS with eight state-of-the-art deep models for paraphrase identification, including DSSM [19], CDSSM [20, 39], MV-LSTM [24], ARC-II [21], MatchPyramid [1], Match-SRNN [23], MP-DOT [1], and uRAE [25]. For DSSM, CDSSM, MV-LSTM, and Match-SRNN, the reported experimental results are provided by [18]. The experimental results of ARC-II, MatchPyramid, MP-DOT, and uRAE come from [1, 21, 22], respectively.

Except for the experimental results having been reported in the existing literature, all the parameters of the baselines and the DPIM-ISS are tuned to optimize the evaluation metrics F1 score on the training corpus and the best parameter settings are used on the testing corpus.

4.2.2. Evaluation Metrics. Followed the previous research, the task of paraphrase identification is formalized as a

classification problem and the accuracy and F1 score are used as the evaluation metrics. Accuracy can be formalized as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \tag{15}$$

where TP is true positive, TN means true negative, FP is false positive, and FN represents false negative.

The $F1$ score is the harmonic mean of precision and recall:

$$F1 = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}}, \tag{16}$$

where the precision and recall are defined as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}. \end{aligned} \tag{17}$$

4.2.3. Word Embedding. Word embedding required in the DPIM-ISS model and ARC-I was all learned based on One Billion Word Benchmark Corpus (http://www.statmt.org/lm-benchmark/) that contains nearly one billion sentences with different English words. We chose CBOW which was provided by gensim [40, 41] as the learning model. The dimension of word embedding was set to 300, the size of context window was set to 5, the lowest word frequency was 5, and the learning rate was 0.0002.

4.2.4. Syntactic Features. We used Stanford's parser (https://nlp.stanford.edu/software/lex-parser.shtml) to get the dependency tree of sentences. The results of parser described the syntactic relationship in a sentence by means of the part of speech and the interword dependency. In our experiment, we only preserved the part-of-speech tags and the word dependency tags. These markers were used as the syntactic features, and we simplified these tags in our experiment. For example, we simplified the tag nmod:including as nmod. Then, only 30 syntactic tags were preserved, shown in Table 2.

4.3. Experimental Results and Analysis. The experimental results are summarized in three parts. In Section 4.3.1, we compare DPIM-ISS to the traditional word-matching approaches, the syntax-similarity approaches, the distributed-representations-of-sentences deep models, and the CNN-based models. We compare the performances of DPIM-ISS with other deep models for paraphrase identification in Section 4.3.2. In Section 4.3.3, we analyze the performance of each substructure in our model.

4.3.1. Comparison with the Word-Matching Approaches, the Syntax-Similarity Approaches, and the Distributed-Representations-of-Sentences Deep Models. The main comparison results of our experiments on MSRP, PAN 2010, and PAN 2012 are summarized in Table 3.

First, we compare the performance of DPIM-ISS with word-matching approaches. We observe that the DPIM-ISS outperforms the Jaccard approach, the Cosine approach, and the METEOR approach on F1 score and accuracy. Comparing DPIM-ISS with METEOR, the experimental results show that DPIM-ISS performs better than the method using only lexical features.

In addition, on PAN 2010 and PAN 2012 datasets, the METEOR approach, which takes the synonym matching into account, is significantly higher than the baselines on accuracy and F1 score. This is closely related to the synonym replacement method used in the construction of PAN datasets.

Then, we analyze the performance of DPIM-ISS and syntax-similarity approaches. The experimental results show that the DPIM-ISS has a significant improvement over the syntax-similarity approach. We also note that the improvement on MSRP datasets is lower than that on PAN 2010 and PAN 2012 datasets. Similarly, compared DPIM-ISS with the method Sentence2Vector and ARC-I, we found that the performance improvements on MSRP are lower than those on PAN 2010 and PAN 2012. We conclude that the performance gap is attributed to the construction methods of the MARP dataset and PAN datasets.

For analyzing the differences on performance, we investigate the three datasets and find two main issues: (1) the syntactic structure on MSRP is more similar than those on PAN datasets. (2) Compared with MSRP, the use of words of PAN are significantly different.

Since the MSRP dataset was constructed using the corpus of topic-clustered news data, it does not adopt the deliberate obfuscation, which results in the small lexical differences but similar syntactic structures between the two sentences in MSRP. Therefore, DPIM-ISS does not get much more benefits than the traditional deep learning methods. For the two PAN datasets, the source sentences are paraphrased in order to avoid plagiarism detection. The vocabulary shows the significant variations, and the syntactic structure takes on the marked difference. By decomposing the sentence's syntactic structure using the dependency tree, we obtain the key substructures of a sentence. The same substructures may be owned by the two sentences simultaneously (such as the predicate verb). Although these substructures present different appearance in terms of words, they may have similar semantics. DPIM-ISS uses the sentence expression interacting the semantics with syntax to obtain the semantic expression on the syntactic structures and learns the patterns of paraphrase in these semantic expressions using CNN. It pays attention to the different functions of semantic matching in different syntactic structures on paraphrase identification and solves the issues of the different syntactic structures as well as the different words to a certain extent.

4.3.2. Comparison with Other Deep Models for Paraphrase Identification. Based on the MSRP dataset, we compare the performance of DPIM-ISS with other main deep models for paraphrase identification. We choose the MSRP dataset

Table 2: Syntactic features.

No.	Feature	No.	Feature	No.	Feature	No.	Feature	No.	Feature
1	advcl	7	JJR	13	RB	19	dobj	25	nsubjpass
2	advmod	8	JJS	14	RBR	20	FW	26	nummod
3	ccomp	9	neg	15	RBS	21	iobj	27	VBG
4	CD	10	NN	16	root	22	JJ	28	VBN
5	csubj	11	NNP	17	VB	23	NNS	29	VBP
6	csubjpass	12	NNPS	18	VBD	24	nsubj	30	VBZ

Table 3: Performance comparisons with word-matching-based approaches, the syntax-similarity approaches, the text-semantic-representation-based deep models, and the CNN-based models.

		MSRP		PAN 2010		PAN 2012	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
Word-matching-based	Jaccard	72.06	81.53	86.26	85.86	53.53	69.73
	Cosine	70.89	81.69	85.23	84.87	65.12	67.45
	METEOR	73.10	81.06	89.50	88.90	82.11	80.70
Syntax-similarity-based	Syntax-sim	66.90	80.03	74.57	72.10	62.74	69.65
Text-semantic-representation-based	Paragraph vector	67.42	80.21	67.33	70.45	51.08	66.48
Deep models	ARC-I	69.60	80.27	50.01	66.68	50.14	66.39
Our model	DPIM-ISS	**73.57**	**83.55**	**91.10**	**91.07**	**83.60**	**82.56**

Table 4: Comparison with other deep models for paraphrase identification on the MSRP dataset.

Deep models	Accuracy	F1
DSSM	70.09	80.96
CDSSM	69.80	80.42
MV-LSTM	75.40	82.80
ARC-II	69.90	80.91
MatchPyramid	**75.94**	83.01
Match-SRNN	74.50	81.70
MP-DOT	75.94	83.01
uRAE	76.8	83.60
DPIM-ISS	73.57	**83.55**

since the results of various deep models for paraphrase identification can be obtained directly from the literature which proposed these models. The data listed in Table 4 come from the experimental results presented in the corresponding literature.

From Table 4, we can see that uRAE and DPIM-ISS, which are built based on the syntactic information, perform much better than the other baselines. Though the best performance of our model (83.55) is still slightly worse than uRAE on F1 score (83.6%) [22], uRAE relies heavily on pretraining on an external large dataset annotated with parse tree information to learn the representation of phrase features for each node in a parse tree. Compared with uRAE, DPIM-ISS only needs to parse the two sentences to be recognized for obtaining the syntactic structures without any additional pretraining.

4.3.3. Model Analysis. First, we analyze the influence of lexical features on DPIM-ISS. We remove the lexical features in DPIM-ISS and use the features captured by the convolutional neural network from the interacting sentence expression as the input of MLP directly to learn the classifier.

Table 5: The effect of lexical features on DPIM-ISS.

Model	MSRP		PAN 2010		PAN 2012	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
DPIM-ISS$_{-L}$	70.50	81.84	86.77	87.74	68.40	72.53
DPIM-ISS	**73.57**	**83.55**	**91.10**	**91.07**	**83.60**	**82.56**

The model that removes the lexical features is denoted as DPIM-ISS-L. Table 5 lists the performance comparison between DPIM-ISS-L and DPIM-ISS.

The experimental results in Table 5 demonstrate that the lexical features help to improve the performance of paraphrase identification, especially on the PAN 2012 dataset. We conclude that METEOR evaluation measures take the synonym replacement into account, which is one of the main construction strategies of the PAN 2012 dataset. However, on the MSRP dataset, there are little changes in the use of words and the syntactic structures, so the additional lexical features do not lead to a significant improvement on MSRP than on PAN 2012.

For the number of syntactic feature parameters, we compared the performance of 30 syntactic features with 67 syntactic features. On MSRP training corpus, we got 0.7119 on accuracy and 0.8173 on F1 when we used 30 syntactic features (the syntactic features in Table 2). However, when we used 67 syntactic features (30 syntactic features in Table 2 added another 37 syntactic features), we got 0.6805 on accuracy and 0.8041 on F1. We also tried two commonly used dimensions of word embedding, 300 and 600, on MSRP training corpus. The accuracy got by the 300 dimensions word embedding was 0.7119 on accuracy and 0.8173 on F1, while the 600-dimensional word embedding achieved 0.6786 on accuracy and 0.7891 on F1. The above two experimental results show that too many features will

affect the classification performance on the size of network that we designed. To further improve the performance of DPIM-ISS, we can try to expand the network size or add the network layers to enhance the representation ability of DPIM-ISS.

5. Related Work

Early work on paraphrase identification usually relied on lexical, semantic, or syntactic similarity measures to identify paraphrases.

Lexical-based approaches used the bag-of-words representations without considering the semantics of the words, which inevitably led to the problem of "polysemy" and "synonymy" in paraphrase identification.

Some methods resorted to the knowledge base (such as WordNet) to measure the word semantic similarity for alleviating the restrictions of word matching-based paraphrase identification methods. For example, Mihalcea et al. utilized the WordNet-based measures to compute the word semantic similarity [8], Mohamed and Oussalah also presented to use the WordNet and Wikipedia to compute the word semantic similarity and named-entity semantic relatedness for paraphrase identification [42], Madnani et al. exploited the METEOR (based on WordNet) machine translation metrics as the features of classifiers to determine the paraphrase [6], and Islam et al. [43] and Bollegala et al. [44] computed semantic similarity using a corpus-based measure. The main advantage of knowledge base-based semantic approaches is that it can make full use of the prior knowledge of experts. However, the limitations of this kind of approaches mainly include the following: knowledge base needs the human maintenance and updating, the limitation of vocabulary coverage, and the lack of sufficient context information to determine the exact concepts.

On the other hand, researchers have noticed the role of syntactic features in paraphrase identification and presented some syntax-based methods. For example, Das and Smith believed that the paraphrase was related to the syntactic structure, and they used the part-of-speech tag and the syntactic dependence of words as the features to learn the classifier [2], Koroutchev et al. exploited the Lempel–Ziv algorithm to compares the syntactic and morphological features of the two texts to detect the text similarity [13], Elhadi and Al-Tobi utilized the part-of-speech sequence to represent text and detect plagiarism [12, 15], Potthast et al. employed n-grams of the syntactic structure sequence to detect the plagiarism in European languages [14], and Mohammad et al. extracted the POS tags as syntactic features of classifiers to identify the paraphrase for the Arabic language [45]. However, these methods could not work effectively when the syntactic structures changed greatly.

To avoid the disadvantages of single class of similarity measures, a different way to look to paraphrase identification is relying on the supervised learning to combine the lexical, syntactic, and semantic features to classify the sentence pair paraphrase or not [46].

In recent years, the distributed representation of words or text has made progress of the semantic representation. Manning pointed out that having a dense, multidimensional representation of similarity between all words was incredibly useful in natural language processing [47]. The distributed representation uses the vectors in contiguous semantic space to project the linguistic units, which makes the similarities of words can be calculated using the distances of word vectors. Thus, two sentences, represented as two vectors in the low-dimensional semantic space, can still have a high similarity even if they do not share any term [39].

Inspired by the success of the deep neural networks recently, the paraphrase identification has been innovated towards the deep paraphrase identification models, including the full-connected neural network-based models such as DSSMs (deep structured semantic models) [19], the CNN-based (convolution neural network) models such as CDSSMs (convolutional deep structured semantic models) [20, 39], ARC-I (Architecture-I) [21], ARC-II (Architecture-II) [21], MatchPyramid [1] and Match-SRNN (Match-special recurrent neural network) [23], the recurrent neural network-based (RNN) models such as MV-LSTM (MV-bidirectional long short-term memory) [24], CNN- and RNN-based models such as Deep-Paraphrase [48], and attention-based alignment models such as pt-DecAtt [49]. These methods focused on the distributed representation of text and identified the paraphrase through the learning of matching degrees and matching patterns, which reduces the dependence on the design of artificial features.

Researchers also introduced the features of syntactic structures into the framework of deep paraphrase identification models. For example, Socher et al. deemed that syntactic and semantic analysis was needed for paraphrase detection, and they presented to exploit recursive autoencoders (RAEs) and unfolding recursive autoencoder (uRAE) to encoder the words, the multiword phrases, and the sentences in syntactic trees [25]. Zhou et al. followed the idea of Socher and used the weighted uRAE to encode the phrases and sentences embedding that obtained from parse trees [50]. Wang et al. proposed the $\text{DeepMatch}_{\text{Tree}}$ to match the two short texts that relied on a tree-mining algorithm [16]. Based on the dependency tree, $\text{DeepMatch}_{\text{Tree}}$ represented the two sentences as the binary matching models composed by the subtree pairs and utilized a deep neural network to learn the matching pattern. Considering the influence of syntactic structure on semantic computation, Liu et al. [51] exploited the syntactic feature for paraphrase identification. In their method, based on the syntactic tree, the TreeLSTM [52] was used to model the sentences and represent the semantic composition. Especially, they introduced the attention mechanism to extract the cross-sentence features. Xu et al. also made use of syntactic features to indicate the dependency relation between words [53]. They incorporated the lexical, syntactic, and sentential encodings for paraphrase identification. In their approach, integrating the syntactic features was verified to contribute to performance improvement. However, the high performance cannot be divorced from the large-scale pretrained model, such as

INPUT : $S = \{(y_{kp}, (s_k, s_p))\}$, iterations
OUTPUT: model
for (sk, sp) in S:
  $(e_{w_1}^{(k)}, e_{w_2}^{(k)}, \ldots, e_{w_i}^{(k)}, \ldots, e_{w_n}^{(k)}) \longleftarrow \text{Embedding}(s_k), (e_{w_1}^{(p)}, e_{w_2}^{(p)}, \ldots, e_{w_i}^{(p)}, \ldots, e_{w_n}^{(p)}) \longleftarrow \text{Embedding}(s_p)$
  $(e_{w_1}^{(k)}, e_{w_2}^{(k)}, \ldots, e_{w_i}^{(k)}, \ldots, e_{w_n}^{(k)}) \longleftarrow \text{SyntaxParsing}(s_k), (g_{w_1}^{(k)}, g_{w_2}^{(k)}, \ldots, g_{w_i}^{(k)}, \ldots, g_{w_n}^{(k)}) \longleftarrow \text{SyntaxParsing}(s_p)$
  for i in 1..n
  $x_{w_i}^{(k)} \longleftarrow e_{w_i}^{(k)} \otimes g_{w_i}^{(k)}$, $x_{w_i}^{(k)} \longleftarrow e_{w_i}^{(k)} \otimes g_{w_i}^{(k)}$
  $T^{(k)} \longleftarrow T^{(k)} + x_{w_i}^{(k)}$, $T^{(p)} \longleftarrow T^{(p)} + x_{w_i}^{(p)}$
  $A_{k,p} \longleftarrow T^{(k)} - T^{(p)}$
  $z_{k,p} \longleftarrow \text{ExtractingLexicalFea}(s_k, s_p)$
  $T \longleftarrow \text{append}(y_{k,p}, A_{k,p}, z_{k,p})$
for iter in range(iterations):
  model $\longleftarrow$ TrainingModel(T)
return model.

Algorithm 1: Training DPIM-ISS.

BERT (bidirectional encoder representations from transformers) [54].

The above approaches enjoyed the advantages of integrating the syntactic features in the paraphrase identification. They all exploited the dependency trees to obtain the local substructures of words or phrases on the syntactic structures at different granularities and learned the semantic representation of these substructures. In this regard, the ideas of this paper are the same as those of the existing work. The difference lies in the semantic representation and interaction on syntactic structures. DPIM-ISS is designed to interact the semantics and syntactic features for obtaining the semantic representation on syntactic structures. Furthermore, we exploit the explicit syntactic structure to model the semantic interaction on syntactic structures between two sentences. This allows us to learn the paraphrase pattern from the semantics on different linguistic features, which was not performed in the RAE, uRAE, weighted uRAE, and $\text{DeepMatch}_{\text{Tree}}$.

6. Conclusions

In this paper, we present the DPIM-ISS, a novel text deep paraphrase identification model interacting semantics with syntax. In DPIM-ISS, we introduce the syntactic information by capturing the syntactic structures and represent the semantics by means of the distributed representation method. Then, we exploit the tensor to interact the semantics and syntax for representing the sentences and use the convolutional neural network to extract the paraphrase patterns in text matching space. Experiments on MSRP, PAN 2010, and PAN 2012 corpus demonstrate that DPIM-ISS achieves comparable or better performance against the traditional word-matching approaches, the syntax-similarity approaches, the distributed-representations-of-sentences-based models, the CNN-based models, and some text deep paraphrase identification methods.

There is an important direction to improve the performance of DPIM-ISS. We note that the acquisition of syntactic features now mainly relies on the results of syntactic parsing. The advantage of this kind of approach is to capture the explicit syntactic structures. However, we can try to another way of exploiting syntactic features, for example, to integrate the representation and the learning of the syntactic features into the network of DPIM-ISS directly. This should be one of our future work.

Appendix

A. Algorithm for Sentence Representation Interacting Semantics with Syntax and Training Process

Sentence representation interacting semantics with syntax and training process is presented in Algorithm 1.

Data Availability

The data of MSRP are available at https://www.microsoft.com/en-us/download/details.aspx?id=52398, the data of PAN 2010 are available at http://bit.ly/mt-para, and the data of PAN 2012 are available at https://pan.webis.de/data.html#pan12-text-alignment. The data used to support the findings of this study are also available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (nos. 61806075 and 61772177).

References

[1] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 2793–2799, Phoenix, AZ, USA, February 2016.

[2] D. Das and N. A. Smith, "Paraphrase identification as probabilistic quasi-synchronous recognition," in *Proceedings of the the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 468–476, Singapore, August 2009.

[3] C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," in *Proceedings of the the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 17–24, New York, NY, USA, June 2006.

[4] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 475–482, Singapore, July 2006.

[5] P. Clough, R. Gaizauskas, S. S. L. Piao, and Y. Wilks, "METER: MEasuring TExt Reuse," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 152–159, Philadelphia, PA, USA, July 2002.

[6] N. Madnani, J. Tetreault, and M. Chodorow, "Re-examining mtranslation metrics for paraphrase identification," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 182–190, Montréal, Canada, June 2012.

[7] H. Li and J. Xu, "Semantic matching in search," *Foundations and Trends in Information Retrieval*, vol. 7, no. 5, pp. 343–469, 2014.

[8] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proceedings of the 21st National Conference on Artificial Intelligence-Volume1*, AAAI Press, Boston, MA, USA, pp. 775–780, 2006.

[9] Y. Zhang and J. Patrick, "Paraphrase identification by text canonicalization," in *Proceedings of the Australasian Language Technology Workshop*, pp. 160–166, Sydney, Australia, 2005.

[10] W. Guo and M. Diab, "Modeling sentences in the latent space," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 864–872, Stroudsburg, PA, USA, 2012.

[11] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 133–149, 2012.

[12] M. Elhadi and A. Al-Tobi, "Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures," in *Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, pp. 679–684, Seoul, Korea, November 2009.

[13] K. Koroutchev and M. Cebrián, "Detecting translations of the same text and data with common source," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 10, Article ID P10009, 2006.

[14] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 45–62, 2011.

[15] M. Elhadi and A. Al-Tobi, "Use of text syntactical structures in detection of document duplicates," in *Proceedings of the third IEEE International Conference on Digital Information Management (ICDIM)*, London, UK, November 2008.

[16] M. Wang, Z. Lu, H. Li, and Q. Liu, "Syntax-based deep matching of short texts," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 1354–1361, Buenos Aires, Argentina, July 2015.

[17] N. Chomsky, "The logical basis of linguistic theory," in *Proceedings of the 9th International Congress of Linguists*, Mouton and Co., Hague, Netherlands, pp. 914–978, 1964.

[18] L. Pang, Y. Lan, J. Xu, J. Guo, S. Wan, and X. Cheng, "A survey on deep text matching," *Chinese Journal of Computers*, vol. 39, no. 126, pp. 985–1003, 2016.

[19] P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, pp. 2333–2338, San Francisco, CA, USA, October 2013.

[20] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 373-374, Seoul, Korea, April 2014.

[21] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proceedings of theAdvances in Neural Information Processing Systems*, pp. 2042–2050, Montréal, Canada, December 2014.

[22] W. Yin and H. Schütze, "MultiGranCNN: an architecture for general matching of text chunks on multiple levels of granularity," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 63–73, Beijing, China, July 2015.

[23] S. Wan, Y. Lan, J. Xu, J. Guo, L. Pang, and X. Cheng, "MatchSRNN: modeling the recursive matching structure with spatial RNN," *Computers & Graphics*, vol. 28, no. 5, pp. 731–745, 2016.

[24] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng, "A deep architecture for semantic matching with multiple positional sentence representations," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 2835–2841, Phoenix, AZ, USA, February 2016.

[25] R. Socher, E. H. Huang, J. Pennin, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pp. 801–809, Granada, Spain, December 2011.

[26] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[27] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing*, pp. 9–16, Jeju Island, Korea, January 2005.

[28] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, Beijing, China, pp. 997–1005, August 2010.

[29] W. Yin and H. Schütze, "Convolutional neural network for paraphrase identification," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 901–911, Denver, CO, USA, May 2015.

[30] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with neural tensor networks for knowledge base completion,"

in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 926–934, Lake Tahoe, USA, 2013.

[31] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *Proceedings of the International Conference on Artificial Intelligence*, pp. 1305–1311, Buenos Aires, Argentina, July 2015.

[32] M. Yu, M. R. Gormley, and M. Dredze, "Combining word embeddings and feature embeddings for fine-grained relation extraction," in *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1374–1379, Denver, CO, USA, May 2015.

[33] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Baltimore, MA, USA, pp. 655–665, June 2014.

[34] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," *The ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization*, vol. 29, pp. 65–72, 2005.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pp. 1097–1105, Lake Tahoe, NV, USA, December 2012.

[36] D. Williams and G. Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–538, 1986.

[37] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *Computer Science*, vol. 3, pp. 1–13, 2015.

[38] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *Computer Science*, vol. 4, pp. 1188–1196, 2014.

[39] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pp. 101–110, Shanghai, China, November 2014.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, AZ, USA, May 2013.

[41] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their Compositionality," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 3111–3119, Lake Tahoe, NV, USA, December 2013.

[42] M. Mohamed and M. Oussalah, "A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 457–485, 2020.

[43] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1–25, 2008.

[44] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 977–990, 2011.

[45] A. L. S. Mohammad, Z. Jaradat, A. L. A. Mahmoud et al., "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features," *Information Processing & Management*, vol. 53, no. 3, pp. 640–652, 2017.

[46] R. Ferreira, G. D. C. Cavalcanti, F. Freitas, R. D. Lins, S. J. Simske, and M. Riss, "Combining sentence similarities measures to identify paraphrases," *Computer Speech & Language*, vol. 47, pp. 59–73, 2018.

[47] C. D. Manning, "Computational linguistics and deep learning," *Computational Linguistics*, vol. 41, no. 4, pp. 699–705, 2015.

[48] B. Agarwal, H. Ramampiaro, H. Langseth, and M. Ruocco, "A deep network model for paraphrase detection in short text messages," *Information Processing & Management*, vol. 54, no. 6, pp. 922–937, 2018.

[49] G. S. Tomar, T. Duque, O. Täckström et al., "Neural paraphrase identification of questions with noisy pretraining," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP, Association for Computational Linguistics 2017*, pp. 142–147, Copenhagen, Denmark, September 2017.

[50] J. Zhou, G. Liu, and H. Sun, "Paraphrase identification based on weighted URAE, unit similarity and context correlation feature," in *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 41–53, Hohhot, China, August 2018.

[51] M. Liu, Y. Zhang, and Y. Chen, "A neural paraphrase identification model based on syntactic structure," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 56, no. 1, pp. 45–52, 2020.

[52] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured Long Short-Term Memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Ling-Uistics and the 7th International Joint Conference on Natural Language Processing*, pp. 1556–1566, Beijing, China, 2015.

[53] S. Xu, X. Shen, F. Fukumoto, J. Li, Y. Suzuki, and H. Nishizaki, "Paraphrase identification with lexical, syntactic and sentential encodings," *Applied Sciences*, vol. 10, no. 12, p. 4144, 2020.

[54] J. Devlin, M. W. Chang, K. Lee et al., "BERT: pre-training of deep bidirectional Transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, MN, USA, 2019.

Hindawi
Complexity
Volume 2020, Article ID 7574609, 14 pages
https://doi.org/10.1155/2020/7574609

Research Article

Multidimensional Heterogeneous Medical Data Push in Intelligent Cloud Collaborative Management System

Gang Liu[1] **and Xiaofeng Li**[2]

[1]*College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China*
[2]*Department of Information Engineering, Heilongjiang International University, Harbin 150025, China*

Correspondence should be addressed to Xiaofeng Li; lixiaofeng@hiu.net.cn

Received 5 July 2020; Revised 3 October 2020; Accepted 7 October 2020; Published 19 October 2020

Academic Editor: Abd E. I.-Baset Hassanien

The medical data in the intelligent cloud collaborative management system have multidimensional heterogeneous interference, and there are problems such as low data information update rate and poor push results in the push process. Therefore, a method for multidimensional heterogeneous medical data push was proposed. First of all, the logical architecture of the multidimensional heterogeneous data push system was determined, and the data push function was designed; secondly, redundant data removal and noise reduction preprocessing were conducted against the push data, correlation rules were used to integrate multidimensional heterogeneous medical data, the weight of medical data was calculated, the heterogeneous data matrix was constructed, and the integrated medical data were weighted to eliminate multidimensional heterogeneous interference. The results show that the data update rate of the proposed method is faster, the user retention and communication rate are high, the data push precision rate is over 80%, and the recall rate is as high as 76%. Therefore, its performance is significantly better than traditional methods.

1. Introduction

Multidimensional medical data refer to medical data with a multidimensional structure defined by dimensions and measures, which is the main object of data online analysis. In recent years, with the development of hospital informatization construction, computer information technology and database management technology in particular have been widely used in the medical field, and a large number of medical data with different dimensions and forms have emerged [1]. As a result, the single data management platform can no longer satisfy the medical data processing requirements [2]. In the information age, cloud computing technology shows obvious advantages and rapid development. To this end, an intelligent cloud collaborative management system with collaborative processing function has been developed for the purpose of coordinating and managing multidimensional heterogeneous medical data, and integrating medical information, and effectively enhancing the application effectiveness of medical data [3, 4]. The main principle of the design and implementation of an intelligent cloud collaborative management system is to combine the task collaboration system and information management system through artificial intelligence technology under the premise of no mutual influence. Therefore, the intelligent cloud collaborative management system can perform work planning tasks and information management tasks, respectively. For hospital management, high-quality data makes management efficiency more effective and more precise [5, 6]. In the management decision-making on medical data, it is usually impossible to quickly extract key data from hospital big data and drive smart healthcare with data, but only through statistical reports submitted by the Statistics Division and Information Division as well as statistical reports in various discrete systems for management decisions [7, 8]; in this case, the intelligent cloud collaborative management system can be analyzed by multidimensional medical data push.

The mass medical information feature push method based on data feature matrix obtained the medical data

feature matrix array by using the medical big data feature intelligent collection method [9, 10], matched all patient group information and part of the patient information by using the data feature matrix array, integrated the corresponding patients into the patient group with the highest similarity, and matched the keywords in the similar patient group and the basic key data features pushed in the medical data feature matrix array, so that the patient group where the patients were located would push medical messages according to the priority push method of location-based service (LBS) [11]. Some scholars analyzed the feasibility of the application of information push technology in the management of tuberculosis patients in the floating population and provided a reference for changing the management model. They used self-made questionnaires to investigate the basic conditions of patients, treatment compliance, personal communication tools, and other information. The conclusion shows that the information push technology is feasible to manage tuberculosis patients in the floating population, and patient acceptance is high [12]. Therefore, the application of information push technology has good feasibility in the management of migrating tuberculosis patients and high patient acceptance. In the proposed method, the Internet-based inpatient health education cloud platform was used to push and set up sharing; that is, based on the characteristics of hospitals and departments [13, 14], this method proposed specific nursing units, regularly pushed propaganda and education information and required its sharing rate to be greater than 80%, regularly shared the feedback information and conclusions, and analyzed the application effect of health propaganda and education cloud platform for Internet-based inpatients according to the patient satisfaction questionnaire, the reading amount of propaganda and education content and the opinions of the patients visited [15].

The above-mentioned traditional management systems can achieve better management functions, and they are widely used in actual work and life. However, according to the studies on their long-term applications, the traditional systems have problems such as long query time and inability to update system information data in time. Therefore, it is necessary to introduce data push technology to improve the updated function of the system. In this system optimization, multidimensional heterogeneous medical data push technology is applied. Among them, "multidimensional" mainly refers to the dimension of data to be pushed, that is, the number of independent parameters in mathematics; "heterogeneous" is a parameter that contains different components and properties. Multidimensional heterogeneous data is the push data message that is not unique in dimension and has a different network structure. Through the application of multidimensional heterogeneous medical data push technology, the function of data push can be realized, thereby solving the problems of long query time and system information data cannot be updated in time in the traditional system and improving the query ability and update speed of the system.

The main contributions of this study are as follows.

(1) The multidimensional heterogeneous medical data push technology is applied to intelligent cloud collaborative management system for analysis; (2) the logical architecture of the multidimensional heterogeneous data push system is determined, which lays the foundation for subsequent research; (3) weighted analysis is made on multidimensional heterogeneous medical data to eliminate multidimensional heterogeneous interference, which provides a basis for improving the efficiency of medical data push; (4) the push channel is selected, which greatly improving the effect of data push.

2. Related Work

By deeply analyzing the Browser/Server (B/S) architecture, Bo [16] designed the corresponding function modules according to the system design principles and equipped with a reasonable database to ensure the quick connection of database; the test results show that the designed electronic medical record management system could quickly enter and query medical record data, achieve functions such as reliable storage of medical record data, and have an important auxiliary role for the hospital to grasp a medical record in time. Jin et al. [17] proposed a wireless intelligent collaboration system based on generalized spatial modulation-media based modulation (GSM-MBM), which could activate multiple antennas at the relay and install radio frequency mirrors near the antennas; different channel paths were constructed by activating different radio frequency mirrors, so as to carry extra information bits; the system transmission efficiency, average paired error probability, and energy consumption gain were deducted according to relevant theories, and Monte Carlo simulations were conducted; as a result, the transmission efficiency was improved, and the bit error rate and energy consumption were reduced; however, the bit error rate was slightly higher under the same transmission efficiency, yet the required the number of required transmitting antennas was greatly reduced, thereby reducing the complexity and cost of system implementation. Aiming at the data characteristics of space structure health monitoring, Zhang et al. [18] put forward the overall framework of the space structure health monitoring Internet of Things (IoT) system and established the application layer data processing algorithm by taking the advantages of cloud computing in processing intensive tasks, completed the design of the cloud data management system for spatial structure monitoring, and conducted real-time processing and interactive display over the monitoring information of multiple large-span spatial structures including National Stadium and Hangzhou Railway Station. Qin et al. [19] designed a medical imaging remote diagnosis cloud service platform to realize automatic uploading, centralized storage and management of image data of primary medical institutions, as well as the sharing of image information and diagnosis reports between hospitals; the system construction and research on the image cloud platform were carried out from the perspectives of registration of image data, the design of data storage center and access to image. Liu et al. [20] proposed a fine-grained access control (FGUR) solution that supported user revocation, which by introducing the attribute hierarchy into the Comparison-Based Encryption

(CBE) and combined with the Broadcast Ciphertext-Policy Attribute-Based Encryption (BCP-ABE), efficiently implemented fine-grained access control and real-time user revocation in the personal health record (PHR) cloud management system; compared with CBE, the FGUR solution shows better performance in encryption overhead and dynamic access permissions.

3. Design of a Collaborative Management System for Multidimensional Heterogeneous Medical Data Push

The multidimensional heterogeneous medical data push technology is the core technology applied in the intelligent cloud collaborative management system. Therefore, the implementation environment and function execution program of multidimensional heterogeneous medical data push technology is introduced for system design.

3.1. Logical Architecture of the System. In the application of multidimensional heterogeneous medical data push technology to the intelligent cloud collaborative management system [21, 22], the server actively sends messages to the receiver, and the system user does not need to actively check and update; the system can push all multidimensional heterogeneous medical data to users via the intelligent cloud server system, the system users can receive the most recent medical data information [23]. Therefore, the logical architecture of the Intelligent Cloud collaboration management system is shown in Figure 1.

According to Figure 1, the logical architecture of the intelligent cloud collaborative management system is mainly composed of a cloud data layer, data management layer, application interface layer, and access layer. Among them, the cloud data layer is to integrate the multidimensional medical data into a data set after receiving the user-level access information. The data management layer is to process the integrated data set to realize the collaborative work of data and push [24, 25]; the application interface layer and access layer are mainly aimed at the receiver, the system can push the required diversified and heterogeneous medical data according to the setting requirements of the user terminal, so as to realize the collaborative work of data and push [26].

3.2. Design of Data Push Function. In the functional design of the collaborative management system software using multidimensional heterogeneous medical data push, this method allows the system to realize the function of medical data push based on the traditional collaborative management function, and tries not to interfere or affect the original collaborative management function during the operation of the new function. Therefore, the push function of multidimensional heterogeneous medical data was specifically designed in this study. It is shown in Figure 2.

According to Figure 2, the heterogeneous data information was mostly collected, and the data was transmitted to the user terminal through the access request; the multidimensional heterogeneous information was collected into the information database, and the weights and push decisions were defined through the access request and then transmitted to the user terminal. After the users successfully subscribed to the content of the cloud push platform, the platform needed to send messages to its own users and push messages to the client in real-time through the long connection established between the cloud and the client [27, 28]. Based on traditional push, the proposed cloud push process was carried out in the cycle of "Subscription-Collection-Decision-Push." The cloud push cycle is shown in Figure 3.

3.3. Cross-Layer Preprocessing of Push Data. In the use of heterogeneous sensors to collect and store original data in the database, this study selects part of the multidimensional heterogeneous medical data in the database as the original information push data. Before the medical data was pushed, cross-layer preprocessing was first performed to reduce the error rate of the data push and to improve the data push quality [29, 30]. The entire data cross-layer preprocessing process is divided into two steps: the removal of redundant data and the noise reduction of data.

Assuming that the original data set is n, the data feature set is $m(f)$, and f represents the eigenvalue, then the probability relationship between the original data set n and the data feature set $m(f)$ can be expressed as follows:

$$D_{\text{sensor}}(f) = n\int_{1}^{R} m(f)d(f), \tag{1}$$

where $D_{\text{sensor}}(f)$ represents the eigenvalue probability of data set n; the solution result of equation (1) is the probability distribution function of the measured value of the medical data push, that is, the push data between the data layer frequency $[1, R]$. According to the solution results, m and n can be divided into three situations. It is shown in Figure 4.

When the solution result is situation 3 in Figure 4, the redundant data in the medical data set n should be removed.

The noise reduction processing was conducted over data set; the frequency-based eigenvalue probability distribution in the multidimensional collaborative processing under the normal transmission link is shown in the following equation:

$$F_x(f) = \iint_{v,\varphi} f(v,\varphi)\mathrm{d}v\mathrm{d}\varphi,\ f(v,\varphi) = u\left(\sqrt{v^2+\varphi^2+\cos\left(\frac{1}{2}\pi+\omega\right)}\right), \tag{2}$$

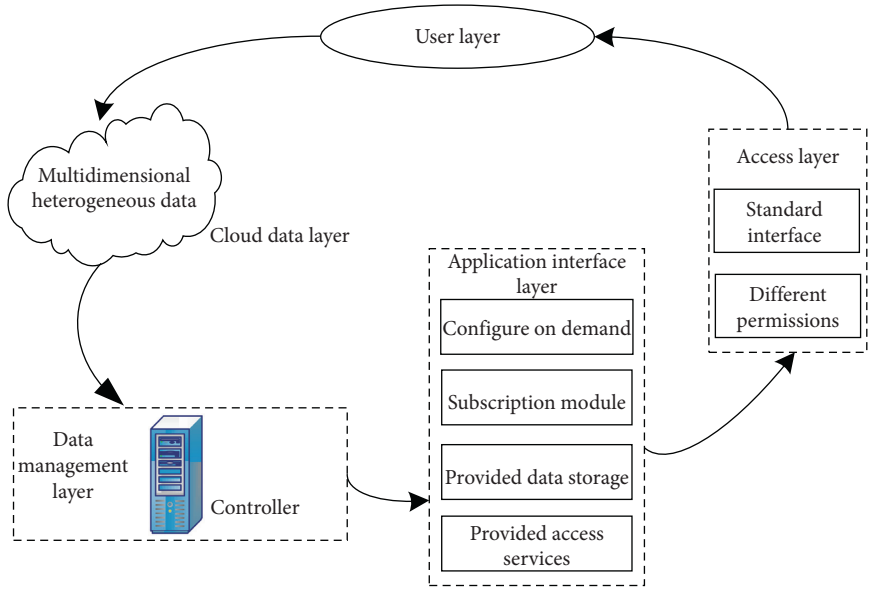


Figure 1: Logical architecture of intelligent cloud collaboration management system.

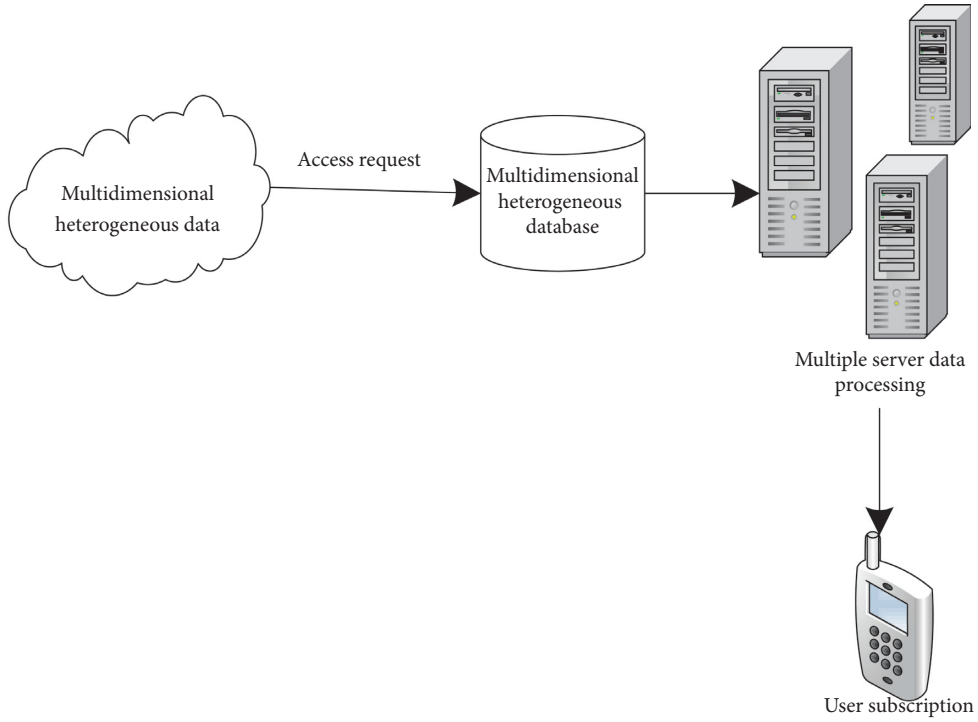


Figure 2: Design of data push function.

where $F_x(f)$ is the eigenvalue probability distribution function under the normal transmission link, and $f(v, \varphi)$ is the probability function. The parameters v and ϕ, respectively, represent the collaboration scale of the push node and the probability of maintaining the collaboration state, and the parameter ω is the angular frequency at which the heterogeneous sensor works.

Enhancement processing is performed on the effective signals in the medical push data, and the enhancement result C is shown in the following equation:

$$C = \begin{cases} F_x(f)\dfrac{v\varphi}{2\omega}, \\ \\ 2v\varphi\cos\omega. \end{cases} \tag{3}$$

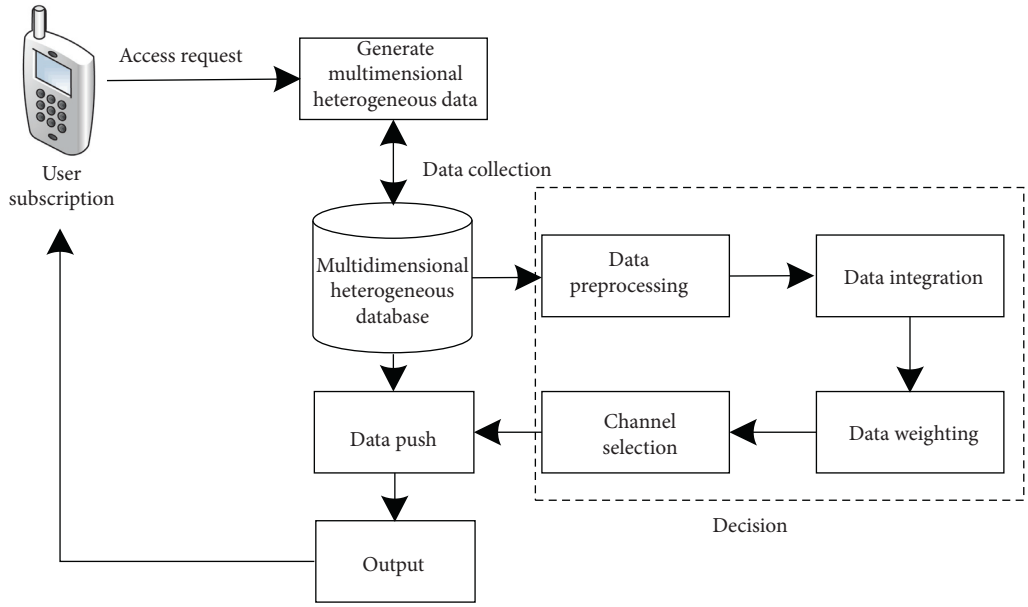


Figure 3: Cloud push cycle.

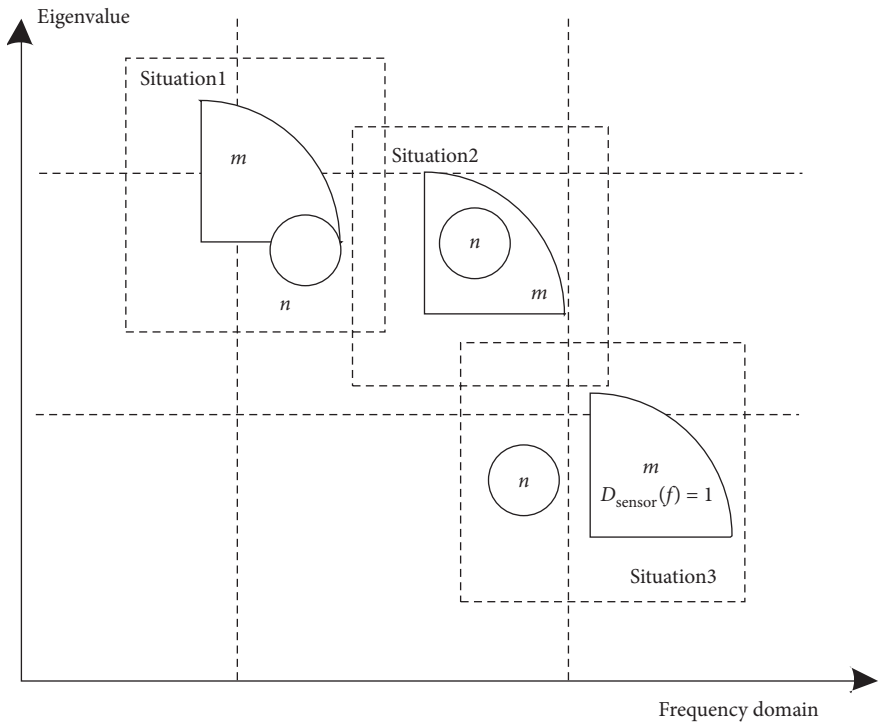


Figure 4: Process of multidimensional heterogeneous data elimination and optimization.

According to equation (3), the strength of the effective data signal can be enhanced while reducing the noise signal [31, 32].

Combined with the above steps, the cross-layer preprocessing of push data is completed.

3.4. Integration of Multidimensional Heterogeneous Medical Data. The principle of integrated management of multidimensional heterogeneous medical data is that the correlation rules algorithm defines strong correlation rule parameters as minimum support and minimum confidence [33]. Among them, the support degree can be specifically defined according to the following equation:

$$\text{support}(A \longrightarrow B) = P(A \cup B). \tag{4}$$

Equation (4) is the probability that multidimensional heterogeneous medical data A and data B appear

simultaneously. If the value of the calculation result is small, the correlation between A and B will be low. Similarly, the confidence in the correlation rule algorithm can be expressed as follows:

$$\text{confidence}(A \longrightarrow B) = P(A|B). \tag{5}$$

Equation (5) is the probability of B appearing when data A appears. If the calculation result is 100%, the correlation between A and B will be high. The correlation of any two data in the sample data can be obtained by synthesizing the two parameters. In addition, the integration results of multidimensional heterogeneous medical data are obtained after the correlated data is clustered and integrated.

3.5. Weighted Analysis of Integrated Medical Data. The analysis of multidimensional heterogeneous medical data is to analyze the content of medical information data. The analysis results can be used as a reference to eliminate multidimensional heterogeneous interference and help the system select an appropriate push method [34]. Besides, the weighted analysis of data was performed after the integration of medical data was completed.

First of all, the weight of the heterogeneous medical data in each dimension was calculated using the following equation:

$$\text{TF}(i) = \sum_{i=1}^{n} tf_j(i) \times class(j), \tag{6}$$

where $\text{TF}(i)$ represents the weight of medical data, $tf(i)$ is the frequency of phrases appearing in a certain area in medical data, and $\text{class}(j)$ is the weight coefficient of regional evaluation, which can be obtained by the system controller [35]. The selected data were arranged in descending order of weight, and the component analysis of medical data was made. The component value is defined by X_i, then the main component value of each multidimensional heterogeneous medical data can be expressed as $X_i(i = 1, 2, \ldots, n)$ and filled into the following equation, so as to get the heterogeneous data matrix:

$$S = \text{TF}(i)[X_1, X_2, X_3, \ldots, X_n], \tag{7}$$

where S is heterogeneous data matrix. The transposed matrix of the matrix S is obtained using the following equation:

$$S^T = \text{TF}(i)\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \cdots \\ X_n \end{bmatrix}, \tag{8}$$

where T represents the transposed symbol. Equations (4) and (5) are multiplied and the average value is obtained as avg_X. Then, the overall heterogeneous value of the multidimensional heterogeneous medical data can be calculated using the following equation:

$$V = S^T \sum \left\{ X_i - \mu \left| \text{avg}_X - \mu \right|^T \right\}, \tag{9}$$

where N is the total amount of initial integrated medical data, and μ is the data offset parameter. If multidimensional heterogeneous medical data needs to be pushed simultaneously, the principal component of each data should be converted to X_{ij}; therefore, the conversion result of matrix S is shown in the following equation:

$$S = \text{TF}(i)\begin{bmatrix} X_{11} & X_{12} & X_{1j} \\ X_{21} & X_{22} & X_{2j} \\ & & \\ X_{i1} & X_{i1} & X_{ij} \end{bmatrix}. \tag{10}$$

The calculation result of V is then expressed in the form of matrix. Through the calculation of V and its diagonal matrix, the weighted calculation eigenvalue W of the data can be solved. It is shown in the following equation:

$$W = |V - kE| = S\left|V \cdot V^{\wedge}\right|, \tag{11}$$

where E represents the identity matrix, and $V^{\wedge}$ parameter represents the diagonal matrix of V, from which the specific value of eigenvalue k can be obtained. k and V together measure the heterogeneity of different medical data.

3.6. Selection of Data Push Channel. After the weighted analysis of integrated medical data, preparations were done to push information to the system. A push channel needed to be utilized in this process. Therefore, it was crucial to select the data push channel, which would directly affect the accuracy of push results.

The selection of data push channel should be specifically considered from two aspects: the carrying capacity of the push channel and the length of the push channel. First, the maximum efficiency of data transmission between a certain user u and the push server r should be calculated. It is shown in the following equation:

$$U_{\text{ur}} = \max_{r \in R}\left[\frac{c_{\text{ur}}}{D_{\text{ur}}}\right], \tag{12}$$

where c_{ur} represents the number of data transmissions used by user u, and D_{ur} represents the total amount of data transmission required by user u. In addition, the matching value MTM_{ur} of the maximum task medical data needs to be calculated:

$$\text{MTM}_{\text{ur}} = U_{\text{ur}}\varphi, \tag{13}$$

where φ is the matching parameter. All the channels that meet the requirements of the above equations are taken as candidate channels, and the channels are arranged according to the transmission distance of the channels. The specific arrangement is shown in Figure 5.

In Figure 5, the signaling channel controls the connection of channels and transfers network management information; physical channel reconfiguration can be used to achieve cofrequency hard handover and compression. The

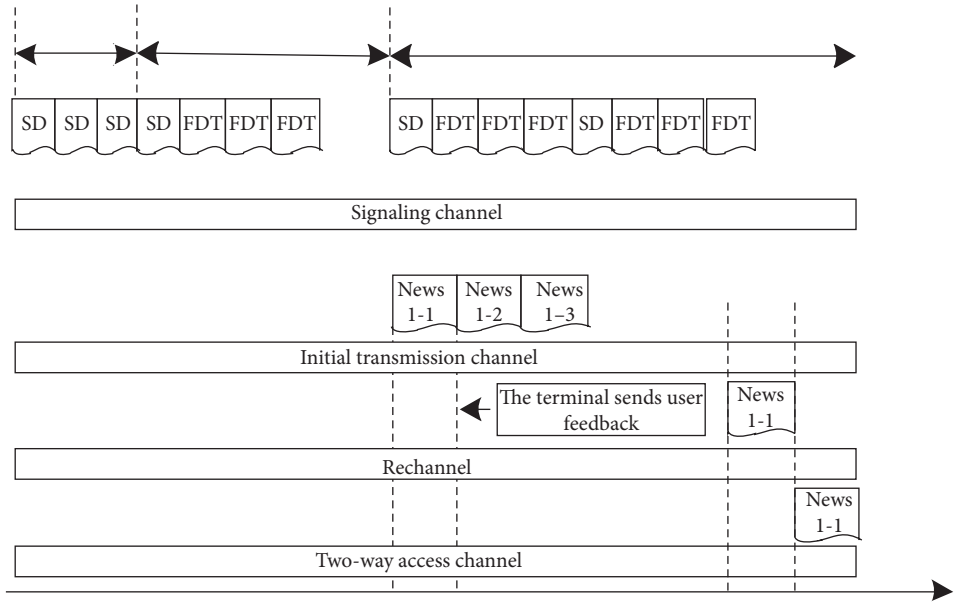


Figure 5: Schematic diagram of push server channel arrangement.

initial transmission channel is responsible for the initial input of medical data, the retransmission channel is responsible for sending commands, and the bidirectional access channel provides diversity gain for wireless transmission and improves the reliability of data transmission. If the channel environment is the same, the value is the same, and the channel that meets the restriction conditions and has the shortest push transmission distance can be selected as the push channel of medical data by calculation.

The optimized multidimensional heterogeneous medical data will be pushed through the selected push channel. Before the data push, it is necessary to perform status processing on the sender and the receiver. Afterwards, the corresponding multidimensional heterogeneous medical data are pushed. During the push process, it is necessary to strictly control the data push quality using the controller. It is shown in the following equation:

$$S_S = \mathrm{MTM}_{\mathrm{ur}} \frac{\nu\varphi}{2\omega} \cdot \frac{p\left(d_{\mathrm{sensor}} \mid t\, n_{oi}\right)}{p\left(e \mid n_{oi}\right)}, \qquad S_{S'} = 2\nu\varphi \cos\omega, \tag{14}$$

where d_{sensor} represents the channel transmission distance, e represents the control parameter, and S_s and $S_{s'}$ represent the data quality eigenvalue pushed by the sender and receiver, respectively. When the values of S_s and $S_{s'}$ are within the interval $[0, 1]$, it is determined that the medical data conforms to the push quality and is passed. When the receiver displays the corresponding push message, the intelligent cloud collaborative management system implements the medical data push function.

3.7. Implementation of Multidimensional Heterogeneous Medical Data Push Management. The management of multidimensional heterogeneous medical data push could be implemented after the selection of the above channels, and the push process is described as follows:

Input: original multidimensional heterogeneous medical data set n and channel selection

Output: push results of multidimensional heterogeneous medical data

The intelligent cloud collaborative management system is initialized and multidimensional heterogeneous medical data push is performed. The specific steps are described as follows:

(1) Replacement mapping of multisource heterogeneous data. Multisource heterogeneous data is to expand the main components of homogeneous data. In order to facilitate mathematical calculations, a permutation matrix P_m is introduced to conduct permutation mapping on the samples obtained on the cloud server, and the result is recorded as y

$$y = \left(y_\alpha, y_b\right)^T = P_m n. \tag{15}$$

The purpose is to gather the same part of the current sample as the isomorphic data in front of the vector, denoted by y_a, and to place the different part behind the vector, denoted by y_b.

(2) The sample structure obtained from the sampling survey does not match the overall composition, and this structural difference can be eliminated and restored by weighting. k different objects are randomly selected from the data set y of permutation mapping as the initial clustering center;

(3) The weight of each attribute in class k is initialized to the same value, that is, the weight of any class $C_{k'}\ (1 \le k' \le k)$ in attribute $A_t\ (1 \le t \le m)$ is $1/m$;

(4) The weighted measure of dissimilarity of any object $n_i \in y$ and class C_i is defined as follows:

$$D(n_i, z_i) = \sum_{i=1}^{N} S_s\left(n_i^t - A_t\right) + \sum_{i=1}^{N} S_{s'}\left(C_i^t - A_t\right). \tag{16}$$

According to the above equation, the dissimilarity between the object and the class centers are calculated, and the data object is divided into the class represented by the cluster center closest to it according to the nearest neighbor principle;

(5) The cluster centers are updated. Among them, the numerical attribute part is obtained by calculating the average value of the objects in the same class, and the subtype attribute part is obtained by calculating the fuzzy class center.

The fuzzy class center of the subtype attribute part is expressed as follows:

$$z_l^c = \left(z_{l,p+1}^c, z_{l,p+2}^c, \ldots, z_{l,p+m}^c\right). \tag{17}$$

(6) The weights of each attribute in the numerical and subtype data parts of each class in the fuzzy class center are calculated, so as to update the information source.

The data set of the fuzzy class center conforms to the high-dimensional distribution. Therefore, the cloud data sample weights can be calculated using this high-dimensional distribution, and the calculation expression of the weight J is as follows:

$$J = \exp\left\{\frac{1}{2}\left|\left(y - z_l^c\right)^T\right| \sum \left(y - z_l^c\right)^{-1}\right\}. \tag{18}$$

The weight J calculated according to the equation (18) can be used to update the information source. After the information source is updated, the equation for pushing heterogeneous data is as follows:

$$\text{push} = J \sum_{i=1}^{N} y_i{}^T. \tag{19}$$

In summary, the multidimensional heterogeneous medical data push is completed, as shown in Figure 6.

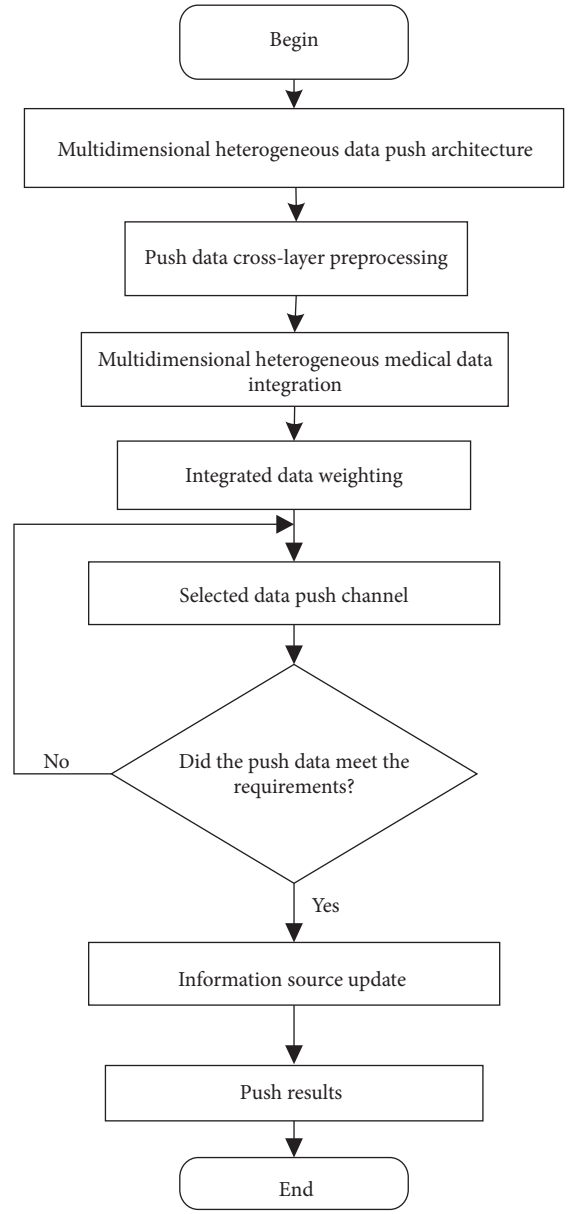


Figure 6: Multidimensional heterogeneous medical data push management process.

4. Experimental Analysis and Results

4.1. Experimental Data. In order to verify the application function of the multidimensional heterogeneous medical data push technology in the intelligent cloud collaborative management system, this study sets up an application analysis experiment. The medical data set is selected as follows:

(1) MIMIC Critical Care Database: the public data set from MIT lab for computational physiology collects data from more than 60000 ICUs, including demographic data, physiological signs, laboratory tests, and drug treatment of patients.

(2) Kent Ridge Biomedical Datasets: database in the biomedical field.

4.2. Experimental Steps. This experiment is carried out in a Matlab environment. The GPU model of Inter® CoreTM i5-4150 is used for cloud collaboration management system training. In this experimental analysis, 10 million data are selected from the above two data sets. The experimental data are processed by the cross-layer preprocessing method in Section 3.3, and effective data are collected and half of the data in each data set is randomly selected as the training

data, and the rest of the data is used as the test data set for experimental analysis. The two datasets of mimic critical care database and Kent ridge biological datasets are used to further optimize the training process of the cloud collaboration management system. The training cycle is set to 100 cycles and the learning rate is set to 0.001.

4.3. Evaluation Criteria

(1) Data update rate: it refers to the data update rate received by the GPS device, which is generally refreshed in units of 1 time/second. Generally, the data update rate determines how much data an instrument can store, and it also represents the speed of data update. Generally speaking, a faster update rate represents better performance.

(2) Retention rate: it refers to the proportion of users who still retain the pushed message within a certain period of time (such as 1–6 weeks), which can also reflect the impact of the model on users to a certain extent.

(3) Communication rate: it refers to the number of communications between the user and the cloud push platform in a unit time, for testing whether the user is willing to use the platform for data push. The higher the communication rate, the stronger the user's willingness to use the system and the better the system performance.

(4) Precision rate of data push:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{20}$$

where FP is the number of samples that are incorrectly pushed. TP is the number of samples correctly pushed.

(5) Recall rate of data push:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{21}$$

where FN is the number of samples incorrectly pushed.

In order to highlight the application value of multidimensional heterogeneous medical data push technology, the architecture method based on B/S in literature [16], the approach based on GSM-MBM in literature [17], the approach based on cloud computing in literature [18], the cloud remote collaboration service system in literature [19], the personal health record cloud management system in literature [20], the priority push based on LBS in literature [21], and the Internet-based inpatient health propaganda and education cloud platform in literature [22] were compared with the proposed method, so as to verify the application value of the proposed method.

4.4. Results and Discussion

4.4.1. Comparison of the Data Update Rate. In this study, the same update task was assigned to different systems. The update task was divided into 7 stages, each update data volume was 1024 MB, and the data update rate of different systems was recorded. It is shown in Figure 7.

According to Figure 7, different collaborative management systems all have higher update rates. After calculation, the average update rates of data for the architecture method based on B/S in literature [16], the approach based on GSM-MBM in literature [17], the approach based on cloud computing in literature [18], the cloud remote collaboration service system in literature [19], the personal health record cloud management system in literature [20], the priority push based on LBS in literature [21], and the Internet-based inpatient health propaganda and education cloud platform in literature [22] are 92.075%, 91.145%, 89.654%, 88.567%, 90.521%, 70%, and 65%, respectively. Due to the application of the multidimensional heterogeneous medical data push technology, the average update data volume of the intelligent cloud collaborative management system is 1021.125 MB, an increase of 77.78 MB, and the average update rate is 99.61%, an average increase of 7.535%. Therefore, it can be concluded that the application of multidimensional heterogeneous medical data push technology can effectively increase the data update rate of the intelligent cloud collaborative management system.

4.4.2. Comparison of the Retention Rate. Actually, the retention rate reflects a conversion rate, that is, the process from the initial unstable users into active users, stable users, and loyal users. With the continuous extension of this retention rate statistical process, the changes of users in different periods can be seen. The higher the retention rate, the better the system performance can meet user needs and the better the performance. It is shown in Figure 8.

According to Figure 8, the average retention rates of the architecture method based on B/S in literature [16], the approach based on GSM-MBM in literature [17], the approach based on cloud computing in literature [18], the cloud remote collaboration service system in literature [19], the personal health record cloud management system in literature [20], the priority push based on LBS in literature [21], and the Internet-based inpatient health propaganda and education cloud platform in literature [22] are 6.5%, 12.5%, 10%, 12.5%, 15%, 10%, and 7.5%, respectively. In contrast, the retention rate of the proposed method is basically stable at 20%, and its retention rate is significantly higher than other methods. Therefore, the proposed method has significant advantages.

4.4.3. Comparison of the Communication Rate. The change in the communication rate of different methods is shown in Figure 9. The communication rates of literature [16], literature [17], literature [18], literature [19], literature [20], literature [21], and literature [22] maintain unchanged between 50% and 65%. In contrast, the communication rate of the proposed method can be as high as about 85%, and finally, tend to stabilize at about 75% as the time of experiment increases. Therefore, the proposed method is much

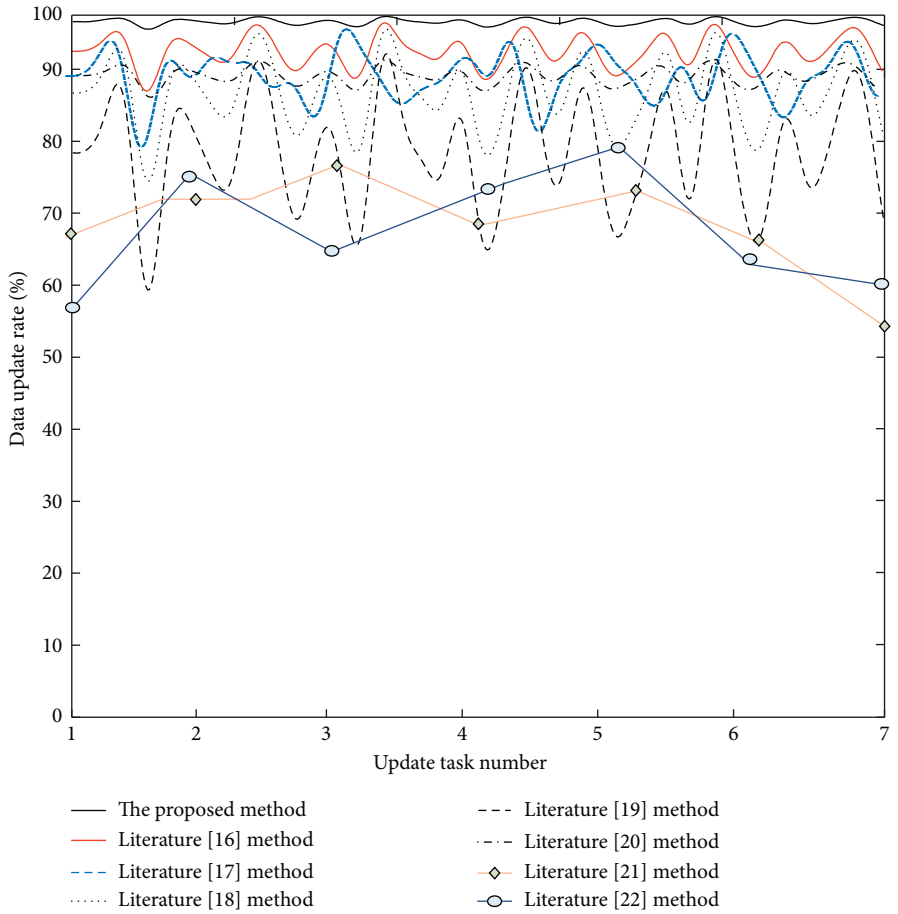


Figure 7: Data update rate change results.

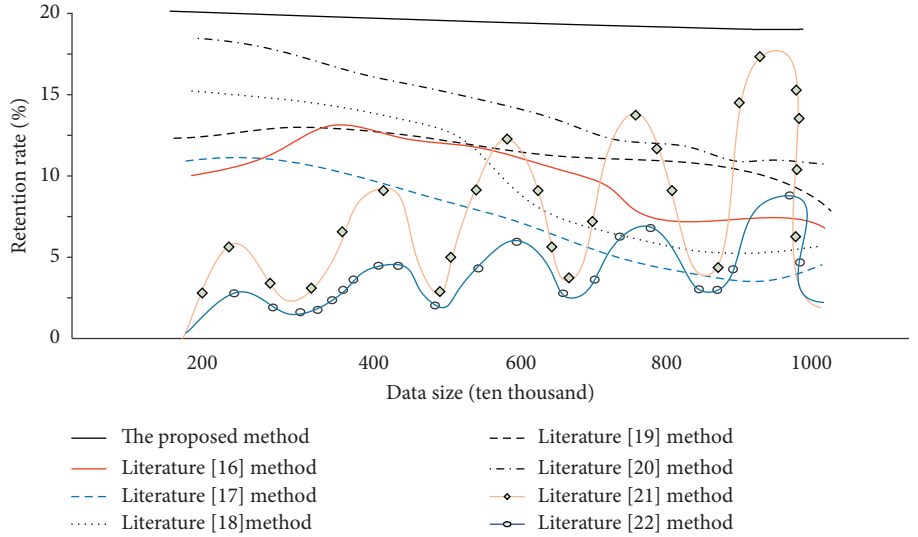


Figure 8: Change of retention rate of different methods.

better than other methods. This is mainly because the proposed method selects the push channel in data push, which increases the accuracy of medical data information push and improves the communication rate.

4.4.4. Comparison of the Precision Rate. The precision rate mainly depends on the specificity of the retrieved information and whether the proposed retrieval strategy can accurately express the users' real intelligence needs. It is shown in Figure 10.

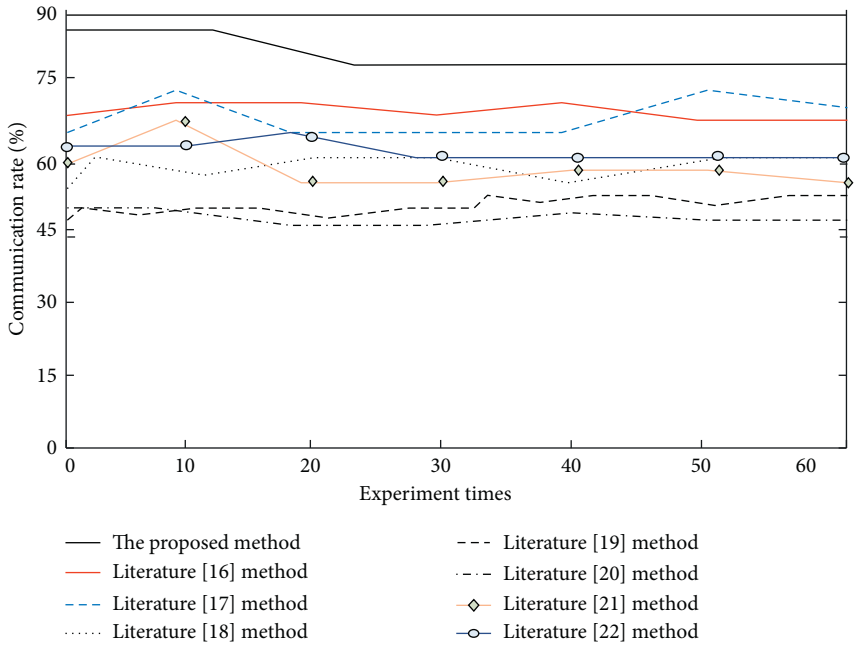


Figure 9: Change of communication rate of different methods.

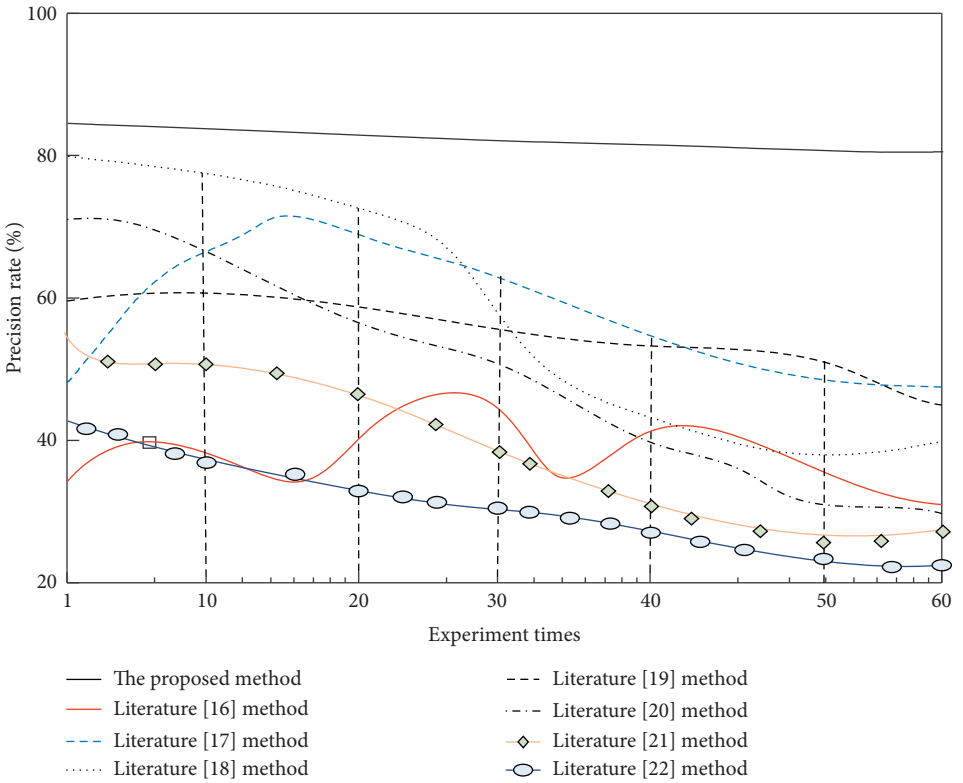


Figure 10: Precision rate result.

According to Figure 10, the average precision rates of the architecture method based on B/S in literature [16], the approach based on GSM-MBM in literature [17], the approach based on cloud computing in literature [18], the cloud remote collaboration service system in literature [19], the personal health record cloud management system in literature [20], the priority push based on LBS in literature [21], and the Internet-based inpatient health propaganda and education cloud platform in literature [22] are 40%, 60%, 60%, 58%, 35%, 30%, and 32%, respectively. In

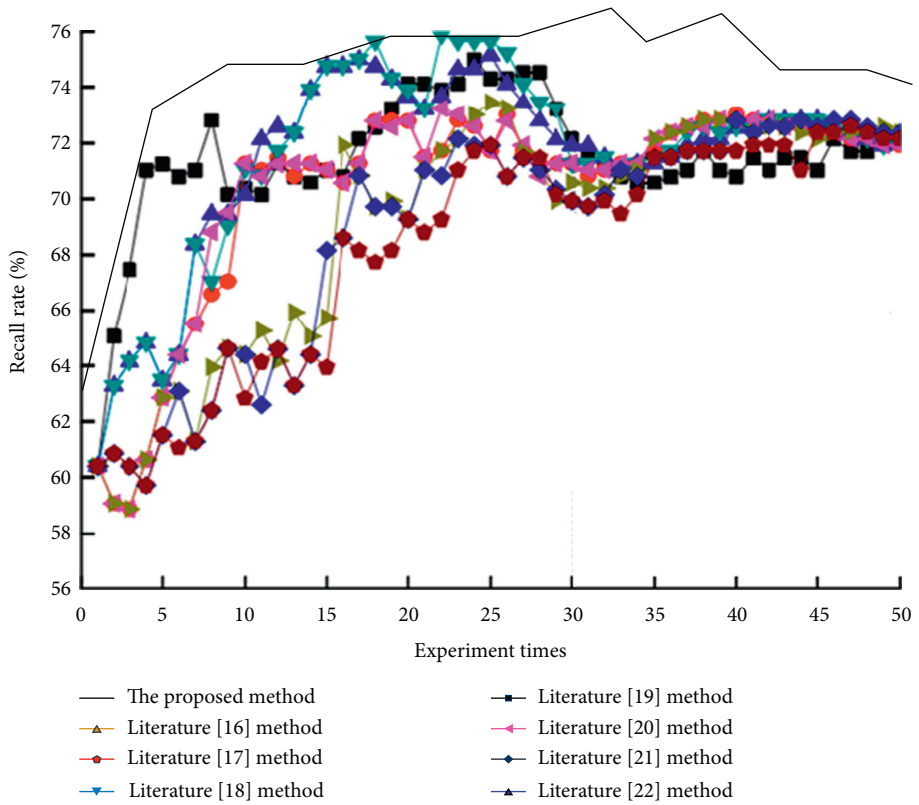


Figure 11: Recall rate results.

contrast, the precision rate of the proposed method is more than 80%. The reason is that the proposed method determines the system logical architecture of multidimensional heterogeneous medical data push, makes a weighted analysis of multidimensional heterogeneous medical data, reduces the dimension of medical data, avoids the removal of multidimensional heterogeneous interference, and improves the precision rate.

4.4.5. Comparison of the Recall Rate. The change in recall rate of different methods is shown in Figure 11. The recall rates of the architecture method based on B/S in literature [16], the approach based on GSM-MBM in literature [17], the approach based on cloud computing in literature [18], the cloud remote collaboration service system in literature [19], the personal health record cloud management system in literature [20], the priority push based on LBS in literature [21], and the Internet-based inpatient health propaganda, and education cloud platform in literature [22] are not uniform, and their entire experimental process is significantly lower than the proposed method. In contrast, the precision rate of the proposed method is up to 76%. The reason is that the proposed method designs a multidimensional cross-layer data preprocessing method, which enhances the strength of data signals and promotes the improvement of the recall rate.

5. Conclusions

The multidimensional heterogeneous medical data is frequently generated in the medical research field, which affects the process of medical data processing. As a result, in-depth research and analysis of multidimensional heterogeneous medical data push is of great benefit to the development of medical systems. This paper analyzes the application of multidimensional heterogeneous medical data push in the intelligent cloud collaborative management system and gives the system logic architecture, and the multidimensional heterogeneous medical data were processed; the push channel was selected, and the data push was effectively completed. The results show that the proposed method has superior performance and good data push performance, which provides a reference basis for the development of the medical field.

However, in actual medical research, due to the diversity of disease types, it is impossible to accurately judge the diseases of many patients, and the accuracy of the results of disease-related data push is affected. Therefore, in further study, we will focus on introducing intelligent systems into disease diagnosis for analysis, providing a data basis for the diagnosis of diverse diseases, helping to obtain more accurate disease diagnosis results, and laying a foundation for further research on the data push system and increasing the intelligence and richness of the system.

Data Availability

The data used to support the findings of this study are included within the article. Readers can access the data supporting the conclusions of the study from MIMIC Critical Care Database and Kent Ridge Biomedical Datasets.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Humanities and Social Sciences Research Planning Fund Project in Ministry of Education under grant number 19YJAZH053, the Opening Project of State Key Laboratory of Digital Publishing Technology under grant number cndplab-2020-M003, and Ministry of Education Science and Technology Development Center Industry-University Research Innovation Fund under grant number 2018A01002.

References

[1] W. Xiao, L. Lu, C. Ji, X. Yu, and D. Qi, "Prediction of water positions in the binding sites of proteins based on collections of multi-source heterogeneous atoms," *Chemical Biology & Drug Design*, vol. 95, no. 2, pp. 224–232, 2020.

[2] Y. Guo, *Design and Implementation of a Collaborative Management Platform Based on Face Verification*, Beijing University of Posts and Telecommunications, Beijing, China, 2018.

[3] P. Sharma, "Prediction of heart disease using 2-tier SVM data mining algorithm," *International Journal of Advanced Research in Big Data Management System*, vol. l, no. 2, pp. 11–24, 2017.

[4] J.-H. Bae and H. Y. Lee, "User health information analysis system of urine and feces separable smart toilet," *International Journal on Human and Smart Device Interaction*, vol. 5, no. 2, pp. 19–24, 2018.

[5] B. Kumwenda, J. A. Cleland, G. J. Prescott, K. Walker, and P. W. Johnston, "Relationship between sociodemographic factors and selection into UK postgraduate medical training programmes: a national cohort study," *British Medical Journal Open*, vol. 8, no. 6, Article ID e021329, 2018.

[6] A. Care, F. A. Ramponi, M. C. Campi et al., "A new classification algorithm with guaranteed sensitivity and specificity for medical applications," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 393–398, 2018.

[7] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan, "Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1317–1332, 2018.

[8] Y. Yuyu, X. Jing, L. Yu, X. Yueshen, X. Wenjian, and Y. Lifeng, "Group-wise itinerary planning in temporary mobile social network," *IEEE Access*, vol. 7, pp. 83682–83693, 2019.

[9] Y. Yin, L. Chen, Y. Xu, and J. Wan, "QoS prediction for service recommendation with deep feature learning in edge computing environment," *Mobile Networks and Applications*, vol. 25, no. 1, pp. 1–11, 2019.

[10] H. Gao, Y. Xu, Y. Yin et al., "Context-aware QoS prediction with neural collaborative filtering for internet-of-things services," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4532–4542, 2020.

[11] K. Jiang, "Research on feature push of massive medical information based on data feature matrix," *Mechanical Design and Manufacturing Engineering*, vol. 48, no. 3, pp. 59–63, 2019.

[12] J. Mei, Y. Wang, J. Zhang et al., "Feasibility analysis of the application of information push technology in tuberculosis management of floating population," *International Medical and Health Guide*, vol. 24, no. 3, pp. 320–323, 2018.

[13] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, p. 1, 2019.

[14] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, In press.

[15] R. Wang, B. Wu, and L. Yan, "Application of Internet-based inpatient health education cloud platform," *International Journal of Nursing*, vol. 38, no. 12, pp. 1758–1761, 2019.

[16] L. Bo, "Design and research of hospital electronic medical record management system based on B.Sc. architecture," *Electronic Design Engineering*, vol. 25, no. 5, pp. 46–49, 2017.

[17] X. Jin, L. Yang, N. Jin et al., "Performance analysis of wireless energy carrying collaboration system based on GSM-MBM," *Journal of Beijing University of Posts and Telecommunications*, vol. 41, no. 5, pp. 141–146, 2018.

[18] S. Zhang, Y. Luo, and Y. Shen, "Design of Internet of things system for health monitoring of spatial structure based on cloud computing," *Spatial Structure*, vol. 23, no. 1, pp. 3–11, 2017.

[19] L. Qin, W. Guo, R. Cai et al., "Construction and practice of medical image cloud remote collaboration service system," *Biomedical Engineering Research*, vol. 27, no. 1, pp. 111–115, 2018.

[20] Q. Liu, X. Liu, B. Hu et al., "Fine-grained access control supporting user revocation in personal health record cloud management system," *Journal of Electronics and Information*, vol. 39, no. 5, pp. 1206–1212, 2017.

[21] J. B. Cole, S. K. Knack, E. R. Karl, G. B. Horton, R. Satpathy, and B. E. Driver, "Human errors and adverse hemodynamic events related to "push dose pressors" in the emergency department," *Journal of Medical Toxicology*, vol. 15, no. 4, pp. 276–286, 2019.

[22] A. F. Cartwright, M. Karunaratne, J. Barr-Walker, N. E. Johns, and U. D. Upadhyay, "Identifying national availability of abortion care and distance from major US cities: systematic online search," *Journal of Medical Internet Research*, vol. 20, no. 5, p. e186, 2018.

[23] B. Fred, "Getting value from EHR data. Analytics push yields payoff at medical center health," *Health Data Management*, vol. 24, no. 3, pp. 51–53, 2016.

[24] M. Gagolewski, R. Perez-Fernandez, and B. De Baets, "An inherent difficulty in the aggregation of multidimensional data," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 3, pp. 602–606, 2020.

[25] M. E. Klijn and J. Hubbuch, "Application of empirical phase diagrams for multidimensional data visualization of high-throughput microbatch crystallization experiments," *Journal of Pharmaceutical Sciences*, vol. 107, no. 8, pp. 2063–2069, 2018.

[26] S. Ali, Al T. Ahmad, Y. Najwa et al., "Data on the relationship between caffeine addiction and stress among Lebanese

medical students in Lebanon," *Data in Brief*, vol. 20, no. 5, p. 104845, 2020.

[27] Y. Jin, X. Guo, Y. Li, J. Xing, and H. Tian, "Towards stabilizing facial landmark detection and tracking via hierarchical filtering: a new method," *Journal of the Franklin Institute*, vol. 357, no. 5, pp. 3019–3037, 2020.

[28] J. Li, X. Zhang, Z. Wang et al., "Dual-band eight-antenna array design for MIMO applications in 5G mobile terminals," *IEEE Access*, vol. 7, no. 1, pp. 71636–71644, 2019.

[29] K. Chen, K. S. Dhindsa, and B. Bhushan, "Collaborative agent-based model for distributed defense against DDoS attacks in ISP networks," *International Journal of Security and Its Applications*, vol. 11, no. 8, pp. 1–12, 2017.

[30] H. Sun and Q. Hu, "A novel deep web data mining algorithm based on multi-agent information system and collaborative correlation rule," *International Journal of Future Generation Communication and Networking*, vol. 9, no. 11, pp. 81–94, 2016.

[31] E. Zhang, J. Fiaidhi, and S. Mohammed, "Social recommendation using graph database Neo4j: mini blog, twitter social network graph case study," *International Journal of Future Generation Communication and Networking*, vol. 10, no. 2, pp. 9–20, 2017.

[32] R. Du, Z. Pei, and J. Tian, "Personalized trusted service recommendation method based on social work," *International Journal of Security and Its Applications*, vol. 10, no. 9, pp. 29–38, 2016.

[33] H. Xi, D. Guo, and H. Zhu, "Application of data mining based on classifier in class label prediction of coal mining data," *International Journal of Security and Its Applications*, vol. 9, no. 10, pp. 425–432, 2015.

[34] M. Francia, M. Golfarelli, and S. Rizzi, "Summarization and visualization of multi-level and multidimensional itemsets," *Information Sciences*, vol. 520, pp. 63–85, 2020.

[35] W. Shao, K. Huang, Z. Han et al., "Integrative analysis of pathological images and multidimensional genomic data for early-stage cancer prognosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 99–110, 2020.

Hindawi
Complexity
Volume 2020, Article ID 3849053, 15 pages
https://doi.org/10.1155/2020/3849053

Research Article

Representation and Reasoning of Three-Dimensional Spatial Relationships Based on R5DOS-Intersection Model Representation and Reasoning Based on R5DOS Model

Jian Li,[1,2] Weijian Zhang,[1] Yating Hu,[1,2] and Zhun Wang[1]

[1]*College of Information Technology, Jilin Agricultural University, Changchun 130118, China*
[2]*Bioinformatics Research Center of Jilin Province, Changchun 130118, China*

Correspondence should be addressed to Yating Hu; huyating@jlau.edu.cn

Received 7 July 2020; Revised 7 September 2020; Accepted 13 September 2020; Published 7 October 2020

Academic Editor: Abd E. I.-Baset Hassanien

This paper aims to disclose the compound topological and directional relationships of three simple regions in the three-dimensional (3D) space. For this purpose, the directional model and the 8-intersection model were coupled into an R5DOS-intersection model and used to represent three simple regions in the 3D space. The matrices represented by the model were found to be complete and mutually exclusive. Then, a self-designed algorithm was adopted to solve the model, yielding 11,038 achievable topological and directional relationships. Compared with the minimum bounding rectangle (MBR) model, the proposed model boasts strong expressive power. Finally, our model was applied to derive the topological and directional relationships between simple regions A and C from the known relationships between simple regions A and B and those between B and C. Based on the results, a compound relationship reasoning table was established for A and C. The research results shed new light on the representation and reasoning of 3D spatial relationships.

1. Introduction

The reasoning of spatial relationship, a.k.a. spatial reasoning, can be implemented quantitatively or qualitatively. Qualitative spatial reasoning, aiming to represent and analyze spatial information, is an important tool in artificial intelligence (AI), machine vision, robot navigation [1, 2], and geographic information system [3].

Over three decades, many theories and models have been developed for spatial reasoning. For instance, Randell et al. [4, 5] put forward the region connection calculus (RCC) theory. Egenhofer and Franzosa [6, 7] proposed the theory of 4-intersection model and 9-intersection model. Li [8] derived a dynamic reasoning method for azimuth relationship.

In recent years, spatial reasoning has evolved rapidly, thanks to the emerging AI applications in image processing [9, 10], computer vision [11, 12], and model prediction [13]. However, most studies on spatial reasoning focus on the spatial relationships on two-dimensional (2D) planes rather than those in three-dimensional (3D) spaces. The 3D space contains too many information elements to be handled by ordinary reasoning methods.

At present, the relationships between objects in the 3D space are mostly solved by compound reasoning. The common approaches of compound reasoning include the compound reasoning of directional and topological relationships [14, 15] and the compound reasoning of directional and distance relationships [16]. Liu et al. [17] designed a 3D improved composite spatial relationship model (3D-ICSRM) in a large-scale environment and proposed a reasoning algorithm to solve that model. The accuracy of the 3D-ICSRM is very limited, and it considers the relationship between qualitative distance and direction. In 2016, Hou et al. [18] extended the convex tractable subalgebra into 3D space and used the BCD algorithm to calculate it. In 2019, Wang et al. [19] extended the oriented point relation algebra (OPRAm) model to 3D and proposed oriented point relation algebra in three-dimensional (OPRA3Dm) algorithm, which

has certain practical significance. These two papers consider the direction relationship. In recent years, the literature mainly studies the relationship between the direction and qualitative distance, while there is less research on the direction and topological relationship. This article will focus on the direction and topological relationship to fill the gaps in this field.

This paper aims to disclose the compound topological and directional relationships of three simple regions in the 3D space. Firstly, the RCC-5 model was combined with a strong directional relationship model for two simple regions, based on the extended 4-intersection theory and spatial orientation relationship in RCC5. The combined model was used to identify the compound topological and azimuth relationships between two simple regions, and solved by a self-designed algorithm. Through programming, a total of 65 topological and directional relationships were obtained in the 3D space.

On this basis, the extended 4-intersection matrix was replaced with an 8-intersection matrix to represent the 3D spatial topological and directional relationships between three simple regions. Then, it was found that the topological and directional relationships between the R5DOS-intersection model of two regions and three regions are complete and mutually exclusive. Further programming reveals a total of 11,038 topological and azimuth relationships between three simple regions in the 3D space and derives a simple topological and directional relationship R (A, C) from two sets of two simple regions R (A, B) and R (B, C).

2. Materials and Methods

2.1. RCC Theory. In 1992, Randell et al. [4, 5] proposed the RCC theory and established the RCC-8 intersection model, which is a boundary-sensitive model. Based on the boundary-sensitive conditions, the RCC-5 intersection model can be derived (Figure 1).

In 1991 and 1995, Egenhofer et al. constructed an extended 4-intersection matrix, which covers two space objects A and B, with A° being the interior of A:

$$\begin{pmatrix} A^o \cap B^o & A^o \cap (B^c)^o \\ (A^o)^o \cap B^o & (A^o)^C \cap (B^c)^o \end{pmatrix}. \tag{1}$$

The value of each position set is either empty or nonempty. Then, the five kinds of relationships in the RCC-5 intersection model can be represented as the matrix in Table 1 and expressed as a set $R_5 = \{(0\ 1\ 1\ 1), (1\ 1\ 1\ 1), (1\ 0\ 1\ 1), (1\ 0\ 0\ 1), (1\ 1\ 0\ 1)\}$.

For three simple areas A, B, and C, $R^2 - \{\partial A \cup \partial B \cup \partial C\}$ can be partitioned into 8 parts (Figure 2).

The eight parts can be illustrated by an 8-intersection matrix:

$$\begin{pmatrix} A^o \cap B^o \cap C^o & A^o \cap B^o \cap (C^c)^o & A^o \cap (B^c)^o \cap C^o & A^o \cap (B^c)^o \cap (C^c)^o \\ (A^c)^o \cap B^o \cap C^o & (A^c)^o \cap B^o \cap (C^c)^o & (A^c)^o \cap (B^c)^o \cap C^o & (A^c)^o \cap (B^c)^o \cap (C^c)^o \end{pmatrix}. \tag{2}$$

The RCC theory fuels the research on spatial relationship models in the past three decades, giving birth to many new theories. Nonetheless, most of these theories target the 2D plane rather than the 3D space. Recently, there is a growing interest in the spatial relationship models of the 3D space, especially the compound reasoning of directional and topological relationships, and that of directional and distance relationships.

2.2. Orientation Model. Minimum bounding rectangle (MBR) is a commonly used model for directional relationship in space [18–20]. The MBR model, 8-direction model, and 16-direction model are shown in Figure 3 below. The MBR model is not consistent with human cognition of directions.

In 2010, He and Bian [21] came up with a special 8-direction cone model (Figure 4), which divides the space into eight regions: NW, NE, EN, ES, SE, SW, WS, and WN. Among them, NW and NE belong to the N direction, EN and ES belong to the E direction, SE and SW belong to the S direction, and WS and WN belong to the W direction.

The 8-direction cone model is easy to describe and recognize and is flexible in dealing with relationships in multiple dimensions. Compared with the 8-direction cone model, the16-direction cone model is also consistent with the human cognition of directions, but too complicated to express. Hence, the 8-direction cone model is more suitable for the reasoning of spatial relationships.

Considering its excellence in spatial segmentation, the 8-direction cone model was coupled with the RCC-5 intersection model for compound reasoning of topological and azimuth relationships in the 3D space.

2.3. Model Construction. Any object in space is wrapped by an outer sphere $\odot A$ with a radius r_A (Figure 5), that is, $\forall (x_A, y_A, z_A) \in \odot A$.

Taking the center of $\odot A$ as the origin of the rectangular coordinate system in space, the spatial Cartesian coordinate system can be established and the reference space can be divided into eight intervals by the x-, y-, and z-axes. Each interval is called a hexagram limit Oct = {1, 2, 3, 4, 5, 6, 7, 8} (Figure 6).

Suppose n points $B_i(x_i, y_i, z_i) \in (i = 1, 2, \ldots, n)$ are scattered in the space. The centroid $b(x_b, y_b, z_b)$ of the point set $B = \{B_1, B_2, \ldots, B_n\}$ can be obtained by k-means clustering (KMC) [20] and treated as the center of the sphere of point set B:

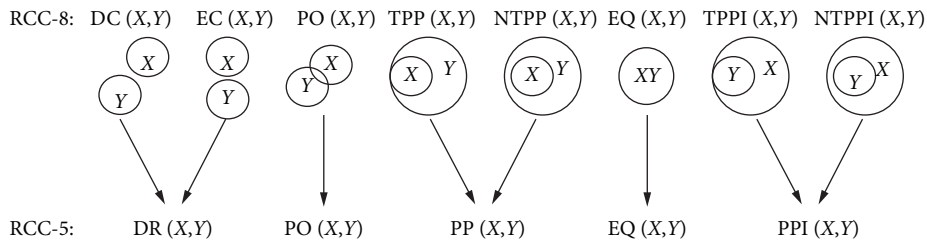


Figure 1: The relationships between RCC-8 and RCC-5 intersection models.

Table 1: Matrix representation of the RCC-5 relationships.

RCC5 relationships	DR (A, B)	PO (A, B)	PP (A, B)	EQ (A, B)	PPI (A, B)
Extended 4-intersection matrix representation	$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

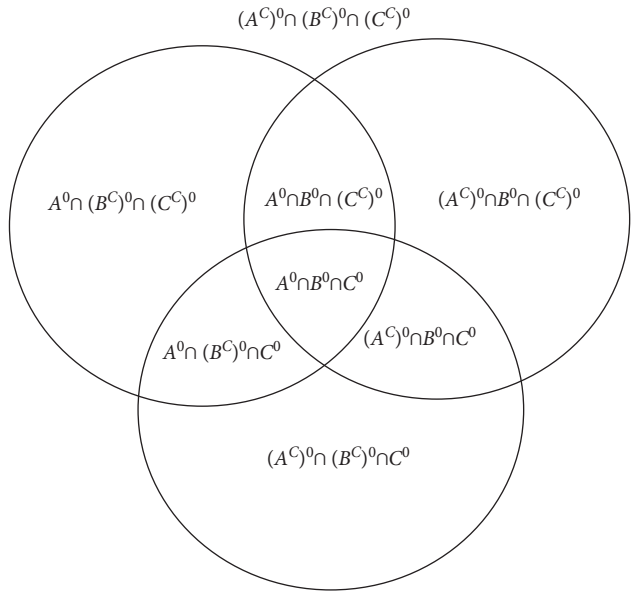


Figure 2: The spatial partition of three simple areas.

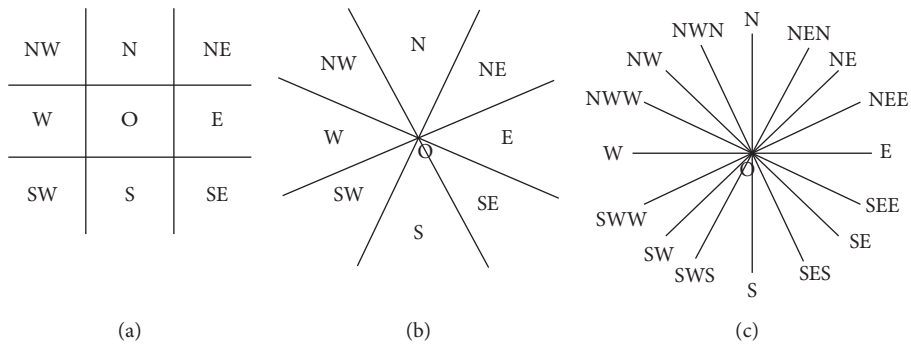


Figure 3: (a) The MBR model, (b) 8-direction model, and (c) 16-direction model.

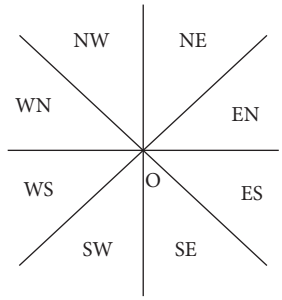


Figure 4: The 8-direction cone model.

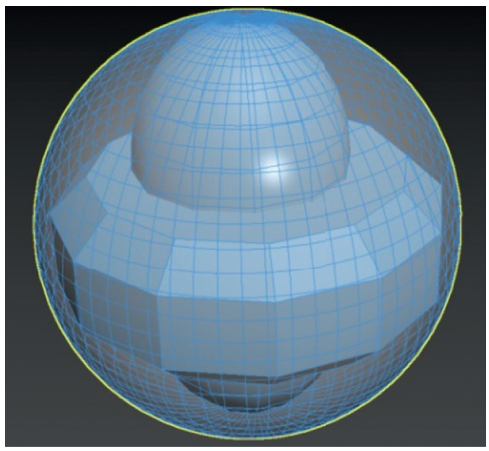


Figure 5: The outer sphere.

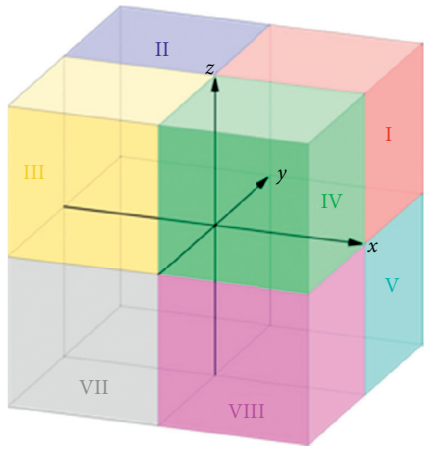


Figure 6: The hexagram limits.

$$\begin{cases} x_b = \dfrac{x_{qB} + x_{pB}}{2}, \\ y_b = \dfrac{y_{qB} + y_{pB}}{2}, \\ z_b = \dfrac{z_{qB} + z_{pB}}{2}. \end{cases} \tag{3}$$

The outer sphere B completely covers the n points: $\forall (x_{Bi}, y_{Bi}, z_{Bi}) \in \odot B$. Similarly, the outer sphere C for point set C can be defined as follows:

$$\begin{cases} x_C = \dfrac{x_{qC} + x_{pC}}{2}, \\ y_C = \dfrac{y_{qC} + y_{pC}}{2}, \quad \forall (x_{Ci}, y_{Ci}, z_{Ci}) \in \odot C, \\ z_C = \dfrac{z_{qC} + z_{pC}}{2}. \end{cases} \tag{4}$$

If it is impossible to find the outer sphere of the space object, the object can be treated as an irregular convex object. Then, five planes π_1: $y = 0, \pi_2$: $x = 0, \pi_3$: $z = 0$, π_4: $y = z$, and π_5: $y = -z$, can be inserted into the rectangular coordinate system in space (Figure 7).

Then, the 3D space can be represented as Dic = {NE, EN, ES, SE, SW, WS, WN, NW}. The angle corresponding to each region can be described as follows:

$$\begin{cases} \theta_{\mathrm{NE}} \in \left[0, \dfrac{\pi}{4}\right), \\ \theta_{\mathrm{EN}} \in \left[\dfrac{\pi}{4}, \dfrac{\pi}{2}\right), \\ \theta_{\mathrm{ES}} \in \left[\dfrac{\pi}{2}, \dfrac{3\pi}{4}\right), \\ \theta_{\mathrm{SE}} \in \left[\dfrac{3\pi}{4}, \pi\right), \\ \theta_{\mathrm{SW}} \in \left[\pi, \dfrac{5\pi}{4}\right), \\ \theta_{\mathrm{WS}} \in \left[\dfrac{5\pi}{4}, \dfrac{3\pi}{2}\right), \\ \theta_{\mathrm{WN}} \in \left[\dfrac{3\pi}{2}, \dfrac{7\pi}{4}\right), \\ \theta_{\mathrm{NW}} \in \left[\dfrac{7\pi}{4}, 2\pi\right), \end{cases} \tag{5}$$

where θ is the dihedral angle of the plane π_i $(i = 1, 2, 3, 4, 5)$. Adding the set of hexagram limits Oct = {1, 2, 3, 4, 5, 6, 7, 8}, the space can be divided into 16 regions:

$$\mathrm{DO} = \begin{pmatrix} 1\mathrm{NE} & 2\mathrm{NE} & 3\mathrm{NW} & 4\mathrm{NW} \\ 1\mathrm{EN} & 2\mathrm{EN} & 3\mathrm{NW} & 4\mathrm{NW} \\ 5\mathrm{ES} & 6\mathrm{ES} & 7\mathrm{WS} & 8\mathrm{WS} \\ 5\mathrm{SE} & 6\mathrm{SE} & 7\mathrm{SW} & 8\mathrm{SW} \end{pmatrix}, \tag{6}$$

where DO is the set of 3D regions and their hexagram limits. If the center of outer sphere B exists in region 1NE, then B strongly exists in that region, denoted as s1NE. If outer sphere B partly exists in region 2NE, then B weakly exists in that region, denoted as w2NE. We let "0" indicate that there is no object in the area, "1" indicates that the object "strongly

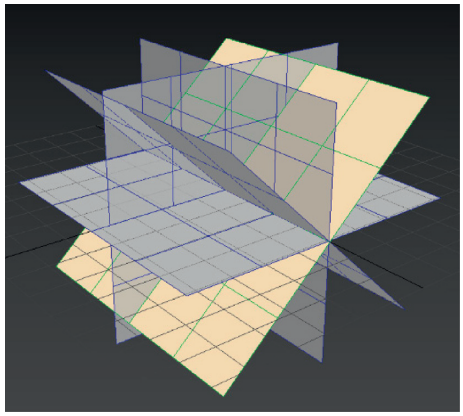
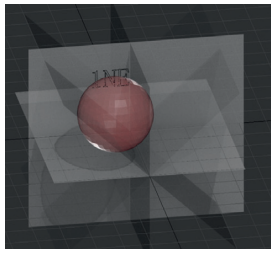


Figure 7: The insertion of five planes.



$$\text{DOS} = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 8: Examples of "weakly presence".

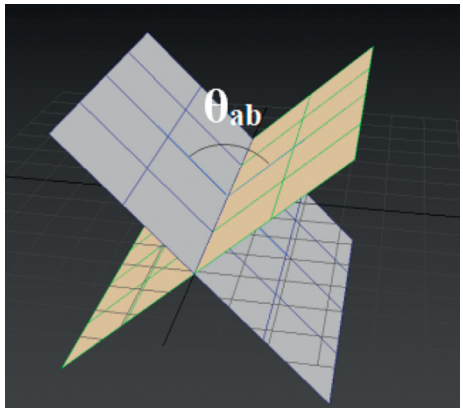


Figure 9: The dihedral angle.

exists" in this area, and "2" indicates that the object "weakly exists" in this area. An example is shown in Figure 8:

$$\text{DOS} = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

For simplicity, only strong existence scenarios were considered. Then, the set of regions, where B strongly exists, can be defined as follows:

$$\text{DOS} = \begin{pmatrix} s1\text{NE} & s2\text{NE} & s3\text{NW} & s4\text{NW} \\ s1\text{EN} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s5\text{ES} & s6\text{ES} & s7\text{WS} & s8\text{WS} \\ s5\text{SE} & s6\text{SE} & s7\text{SW} & s8\text{SW} \end{pmatrix}, \tag{8}$$

where

$$\begin{cases} s1\text{NE}; x_b \geq 0, y_b \geq 0, z_b \geq 0, & \theta_{ob} \in \left[0, \frac{\pi}{4}\right), \\ s2\text{NE}; x_b < 0, y_b \geq 0, z_b \geq 0, & \theta_{ob} \in \left[0, \frac{\pi}{4}\right), \\ s1\text{EN}; x_b \geq 0, y_b \geq 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right), \\ s2\text{EN}; x_b < 0, y_b \geq 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right), \\ s5\text{ES}; x_b \geq 0, y_b \geq 0, z_b < 0, & \theta_{ob} \in \left[\frac{\pi}{2}, \frac{3\pi}{4}\right), \\ s6\text{ES}; x_b < 0, y_b \geq 0, z_b < 0, & \theta_{ob} \in \left[\frac{\pi}{2}, \frac{3\pi}{4}\right), \\ s5\text{SE}; x_b \geq 0, y_b \geq 0, z_b < 0, & \theta_{ob} \in \left[\frac{3\pi}{4}, \pi\right), \\ s6\text{SE}; x_b < 0, y_b \geq 0, z_b < 0, & \theta_{ob} \in \left[\frac{3\pi}{4}, \pi\right), \\ s8SW; x_b \geq 0, y_b \geq 0, z_b < 0, & \theta_{ob} \in \left[\pi, \frac{5\pi}{4}\right), \\ s7SW; x_b \geq 0, y_b \geq 0, z_b \geq 0, & \theta_{ob} \in \left[\pi, \frac{5\pi}{4}\right), \\ s8\text{WS}; x_b \geq 0, y_b < 0, z_b < 0, & \theta_{ob} \in \left[\frac{5\pi}{4}, \frac{3\pi}{2}\right), \\ s7\text{WS}; x_b < 0, y_b < 0, z_b < 0, & \theta_{ob} \in \left[\frac{5\pi}{4}, \frac{3\pi}{2}\right), \\ s4\text{WN}; x_b \geq 0, y_b < 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{3\pi}{2}, \frac{7\pi}{4}\right), \\ s3\text{WN}; x_b < 0, y_b < 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{3\pi}{2}, \frac{7\pi}{4}\right), \\ s4\text{NW}; x_b \geq 0, y_b < 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{7\pi}{4}, 2\pi\right), \\ s3\text{NW}; x_b < 0, y_b < 0, z_b \geq 0, & \theta_{ob} \in \left[\frac{7\pi}{4}, 2\pi\right), \end{cases} \tag{9}$$

where θ_{ob} the dihedral angle formed by planes π_{ob} and π_1, which is perpendicular to the x-axis and passes the straight line ab (Figure 9).

For two regions, the extended 4-intersection matrix can be introduced to the DOS:

$$R5_2\text{DOS} = \begin{pmatrix} A^o \cap B^o & A^o \cap (B^c)^o & (A^c)^o \cap B^o & (A^c)^o \cap (B^c)^o \\ s1\text{NE} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s1\text{EN} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s5\text{ES} & s6\text{ES} & s7\text{WS} & s8\text{WS} \\ s5\text{SE} & s6\text{SE} & s7\text{SW} & s8\text{SW} \end{pmatrix}. \tag{10}$$

For three regions, the 8-intersection matrix can be introduced to the DOS:

$$R5_3\text{DOS} = \begin{pmatrix} A^o \cap B^o \cap C^o & A^o \cap B^o \cap (C^c)^o & A^o \cap (B^c)^o \cap C^o & A^o \cap (B^c)^o \cap (C^c)^o \\ (A^c)^o \cap B^o \cap C^o & (A^c)^o \cap B^o \cap (C^c)^o & (A^c)^o \cap (B^c)^o \cap C^o & (A^c)^o \cap (B^c)^o \cap (C^c)^o \\ s1\text{NE} & s2\text{NE} & s3\text{NW} & s4\text{NW} \\ s1\text{EN} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s5\text{ES} & s6\text{ES} & s7\text{WS} & s8\text{WS} \\ s5\text{SE} & s6\text{SE} & s7\text{SW} & s8\text{SW} \end{pmatrix}. \tag{11}$$

Our model consists of two layers: the first layer is the topological relationship R_5 layer, and the second layer is the orientation relationship DOS layer. Then, the following definition can be derived.

Definition 1. For the orientation relationship layer, there is

$$\varepsilon(\text{DOS}) = \begin{cases} 1, & \text{the enter of the outer sphere } B \text{ exists in this region,} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Suppose

$$\varepsilon\left(R5_2\text{DOS}\right) = \begin{pmatrix} \varepsilon(A^o \cap B^o) & \varepsilon(A^o \cap (B^c)^o) & \varepsilon((A^c)^o \cap B^o) & \varepsilon((A^c)^o \cap (B^c)^o) \\ \varepsilon(s1NE) & \varepsilon(s2EN) & \varepsilon(s3NW) & \varepsilon(s4NW) \\ \varepsilon(s1EN) & \varepsilon(s2EN) & \varepsilon(s3NW) & \varepsilon(s4NW) \\ \varepsilon(s5ES) & \varepsilon(s6ES) & \varepsilon(s7WS) & \varepsilon(s8WS) \\ \varepsilon(s5SE) & \varepsilon(s6SE) & \varepsilon(s7SW) & \varepsilon(s8SW) \end{pmatrix}. \tag{13}$$

For any two simple regions A and B, it is possible to obtain a 5×4‘0-1 matrix. In theory, a total of 2^{20} matrices could be acquired, which correspond to 2^{20} topological and directional relationships in the 3D space:

$$R5_3\text{DOS} = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) & \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap (C^c)^o) \\ \varepsilon(s1\text{NE}) & \varepsilon(s2\text{NE}) & \varepsilon(s3\text{NW}) & \varepsilon(s4\text{NW}) \\ \varepsilon(s1\text{EN}) & \varepsilon(s2\text{EN}) & \varepsilon(s3\text{NW}) & \varepsilon(s4\text{NW}) \\ \varepsilon(s5\text{ES}) & \varepsilon(s6\text{ES}) & \varepsilon(s7\text{WS}) & \varepsilon(s8\text{WS}) \\ \varepsilon(s5\text{SE}) & \varepsilon(s6\text{SE}) & \varepsilon(s7\text{SW}) & \varepsilon(s8\text{SW}) \end{pmatrix}. \tag{14}$$

Based on the topological relationship between outer spheres B and C, the existence of the centers of the two spheres can be described in two cases.

Case 1. Only the center of one outer sphere exists in the current region:

$$\varepsilon(\mathrm{DOS}) = \begin{cases} 0, & \text{no center exists in the current region,} \\ 1, & \text{the center of outer sphere } B \text{ exists in the current region,} \\ 2, & \text{the center of outer sphere } C \text{ exists in the current region.} \end{cases} \tag{15}$$

Case 2. The centers of both outer spheres exist in the current region:

$$\varepsilon(\mathrm{DOS}) = \begin{cases} 0, & \text{no center exists in the current region,} \\ 1, & \text{the centers of both outer spheres exist in the current region.} \end{cases} \tag{16}$$

According to the above conditions, $2^8 \times 3^{16}$ matrices could be obtained theoretically, which correspond to $2^8 \times 3^{16}$ topological and directional relationships in the 3D space.

2.4. Model Properties

Definition 2. In layer R_5, any $m \times n$ -order 0-1 matrices $A = (a_{ij})_{m\times n}$ and $B = (b_{ij})_{m\times n}$ can be defined as $A \cup B = (a_{ij} \vee b_{ij})_{m\times n}$. Then, a 0-1 diagonal matrix can be established as Table 2.

The following proposition can be derived from Table 2:

Proposition 1. $\varepsilon(A \cup B) = \varepsilon(A) \vee \varepsilon(B)$.

For $R_5 = \{(0\ 1\ 1\ 1), (1\ 1\ 1\ 1), (1\ 0\ 1\ 1), (1\ 0\ 0\ 1), (1\ 1\ 0\ 1)\}$, *R (A, B) is the element that corresponds to the topological relationship* R_5 *between any two simple regions A and B.*

Then, the following theorem can be obtained.

Theorem 1. *For simple regions A, B, and C, there exists* $\begin{pmatrix} \varepsilon(A^o \cap B^o) & \varepsilon(A^o \cap (B^c)^o) \\ \varepsilon((A^c)^o \cap B^o) & \varepsilon((A^o)^C \cap (B^c)^o) \end{pmatrix} = R(A, B) \in R_5$.

Similarly, there exists $\begin{pmatrix} \varepsilon(A^o \cap C^o) & \varepsilon(A^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap C^o) & \varepsilon((A^o)^C \cap (C^c)^o) \end{pmatrix} = R(A, C) \in R_5$.

Theorem 2. *In the 3D space given by the R5DOS-intersection matrix, the topological relationship between the three simple regions is mutually exclusive and complete. The DOS space of the R5DOS-intersection matrix, which consists of 16 regions, is a half-open, half-closed interval with mutual exclusion. That is, for any three simple regions A, B, and C in the 3D space, there exists only one relationship satisfied by the ordered pair <A, B, C>.*

Proof. For any three simple regions A, B, and C, the 8 and 16 regions divided by the 8-interesection matrix are disjoint. The 0-1 matrix of three simple regions uniquely corresponds to the matrix derived from the R5DOS-intersection model. In other words, the three regions have the relationship represented by this matrix so that the R5DOS-intersection matrix model gives a complete topological relationship in the 3D space.

Then, it is assumed that the topological relationship between A, B, and C corresponds to two matrices $R5_3$DOSa and $R5_3$DOSb and can be induced by the $R5_3$DOS-intersection model. Then, there exists $1 \le i \le 24$ such that $R5_3\mathrm{DO}_S a_i \neq R5_3\mathrm{DO}_S b_i$. If $i = 1$, $R5_3\mathrm{DOSa}_i = 0$ and $R5_3\mathrm{DOSb}_i = 1$, $A^\circ \cap B^\circ \cap C^\circ$ is both empty and nonempty, which is obviously contradictory. Hence, the above theorem was proved valid.

2.5. Constraints on Two Simple Regions.

Theoretically, two simple regions might correspond to 2^{20} matrices, but there must be 0-1 matrices that cannot be realized. Therefore, the following constraints were designed on two simple regions.

Constraint 1: to correspond to a real-world topological relationship, the 0-1 matrix of layer R_5 must belong to one of the five cases: $R_5 = \{(0\ 1\ 1\ 1), (1\ 1\ 1\ 1), (1\ 0\ 1\ 1), (1\ 0\ 0\ 1), (1\ 1\ 0\ 1)\}$.

Constraint 2: if layer R_5 satisfies $R5 = (1\ 0\ 0\ 1)$, that is, outer spheres A and B are equal, then the 0-1 matrix of the DOS layer is $\mathrm{DOS} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. This means, when outer spheres A and B are equal, the center b of outer sphere B is G, which coincides with that of outer sphere A: $b(x_b, y_b, z_b) = (0, 0, 0)$.

Constraint 3: since the 16 regions are disjoint, they must be mutually exclusive and complete. If R (A, B) does not fall at the center of outer sphere B, it can only exist in one of these regions. In the DOS layer, an outer sphere only exists in one of the 16 intervals. In this way,

Table 2: The 0-1 diagonal matrix.

∨	0	1
0	0	1
1	1	1

a total of 65 directional and topological relationships can be obtained (Table 3).

Case 1

$$\begin{Bmatrix} \text{DRs1NE, DRs2NE, DRs1EN, DRs2EN, DRs5ES, DRs6ES, DRs5SE,} \\ \text{DRs6SE, DRs8SW, DRs7SW, DRs8WS, DRs7WS, DRs4WN, DRs3WN, DRs4NW, DRs3NW} \end{Bmatrix} \in \text{DRall},$$
$$\begin{Bmatrix} \text{POs1NE, POs2NE, POs1EN, POs2EN, POs5ES, POs6ES, POs5SE,} \\ \text{POs6SE, POs8SW, POs7SW, POs8WS, POs7WS, POs4WN, POs3WN, POs4NW, POs3NW} \end{Bmatrix} \in \text{POall},$$
$$\begin{Bmatrix} \text{PPs1NE, PPs2NE, PPs1EN, PPs2EN, PPs5ES, PPs6ES,} \\ \text{PPs5SE, PPs6SE, PPs8SW, PPs7SW, PPs8WS, PPs7WS, PPs4WN, PPs3WN, PPs4NW, PPs3NW} \end{Bmatrix} \in \text{PPall}, \tag{17}$$
$$\begin{Bmatrix} \text{PPIs1NE, PPIs2NE, PPIs1EN, PPIs2EN, PPIs5ES, PPIs6ES,} \\ \text{PPIs5SE, PPIs6SE, PPIs8SW, PPIs7SW, PPIs8WS, PPIs7WS,} \\ \text{PPIs4WN, PPIs3WN, PPIs4NW, PPIs3NW} \end{Bmatrix} \in \text{PPIall}.$$

Case 2

$$\{\text{DRall, POall, PPall, EQ, PPIall}\} \in \Phi. \tag{18}$$

2.6. Constraints on Three Simple Regions. The following constraints were designed on three simple regions.

Constraint 1: to uniquely correspond to the topological and directional relationships in the 3D space, a $R5_3$DOS matrix must satisfy the following conditions.

Definition 3

$$R(A,B) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap C^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon(A^o \cap B^o \cap (C^c)^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \\ (\varepsilon(A^c)^o \cap B^o \cap (C^c)^o) & \left(\varepsilon\left(A^C\right)^o \cap (B^c)^o \cap (C^c)^o\right) \end{pmatrix},$$
$$\begin{pmatrix} \varepsilon(A^o \cap B^o) & \varepsilon(A^o \cap (B^c)^o) \\ (\varepsilon(A^c)^o \cap B^o) & \left(\varepsilon(A^o)^C \cap (B^c)^o\right) \end{pmatrix} \in R_5,$$
$$R(A,C) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (C^c)^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix}, \tag{19}$$
$$\begin{pmatrix} \varepsilon(A^o \cap C^o) & \varepsilon(A^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap C^o) & \varepsilon\left((A^o)^C \cap (C^c)^o\right) \end{pmatrix} \in R_5,$$
$$R(B,C) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) \\ \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon\left(\left(A^C\right)^o \cap (B^c)^o \cap (C^c)^o\right) \end{pmatrix},$$
$$\begin{pmatrix} \varepsilon(B^o \cap C^o) & \varepsilon(B^o \cap (C^c)^o) \\ \varepsilon((B^c)^o \cap C^o) & \varepsilon\left((B^o)^C \cap (C^c)^o\right) \end{pmatrix} \in R_5.$$

(1) $R5_3$DOSaALL ⟵ $2^8 * 3^{16}$ basic topological relationships//All basic topological relationships
(2) $R5_3$DOSa ⟵ null//TR empty Test
(3) for each x in $R5_3$DOSaALL
(4) if x satisfies Constraint 1//if t satisfies Constraint 1Test
(5) if x satisfies Constraint 2//if t satisfies Constraint 2Test
(6) if x satisfies Constraint 3//if t satisfies Constraint 3
(7) $R5_3$DOSa ⟵ {$R5_3$DOSa, x}//If the constraint is satisfied, t is placed in TR
(8) end if
(9) end if
(10) end for
(11) return $R5_3$DOSa//Return result

Algorithm 1

Constraint 2: since all three simple regions are bounded, $(A^C)^\circ \cap (B^C)^\circ \cap (C^C)^\circ$ is always 1.

From Constraints 1 and 2, it can be inferred that layer R_5 has109 topological relationships for any three simple regions in the 3D space.

Constraint 3: after adding the orientation relationship, some topological relationships are not satisfied in the orientation regions. In some topological relationships, the center of an outer sphere will change with that of the other outer spheres. For instance, if layer $R_5 = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$, the outer sphere B will change with the overlap between outer spheres A and C (Figure 10).

Case 1: if the A, B, and C are equal, they can be regarded as one area:

$$\mathrm{DOS} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{20}$$

Case 2: if any two of the three simple regions are equal, the ternary region can be regarded as a binary region with only one 1 in the DOS layer.

Case 3: if any two of the three simple regions are inclusive or noninclusive, the ternary region can be regarded as a binary region when any two regions intersect and the sum of layer R_5 is 4.

Case 4: if only one of the three simple regions is inclusive or noninclusive, the ternary region can be regarded as a binary region when any two regions intersect and the sum of layer R_5 is 5.

Case 5: if the three simple regions are disjoint, the ternary region can be regarded as a binary region when any two regions intersect and the sum of layer R_5 is 5.

Case 6: if simple regions B and C are inclusive or noninclusive and separated from A, then the center of the A can only fall within B and C:

$$(x_{\mathrm{in}}, y_{\mathrm{in}}, z_{\mathrm{in}}) \in P_{\mathrm{out}}. \tag{21}$$

For a ternary reference object in the 3D space, there are theoretically $2^8 \times 3^{16}$ matrices. Under the above constraints, a total of 11,038 matrices were obtained after removing the nonexistent scenarios.

2.7. Topological Relationship Algorithm for 3 Simple Regions in the 3D Space. The topological relationship algorithm for 3 simple regions in the 3D space can be implemented in the following steps.

Step 1: assign each object a row vector [$a1$, $a2$, . . ., $a24$]. Generate a theoretical object of the type $2^8 \times 3^{16}$, i.e., a matrix A of $2^8 \times 3^{16}$ row vectors.

Step 2: scan each row of matrix A, and mark all row vectors that satisfy the constraints.

Step 3: save all the marked row vectors as a matrix B and output the matrix as the final result.

The pseudocode of the algorithm is displayed as follows.
Topological and directional relationship:
Gen (null; $R5_3$DOSa)//Input: null; output: topological relationship satisfying constraints (Algorithm 1).

3. Results and Discussion

3.1. Comparison between $R5_3$DOS-Intersection Model and MBR Model. This section proves that the $R5_3$DOS-intersection model has stronger expressive power than the MBR model in the 3D space [21–23].

First, layer R_5 was defined as $R(A, B)$ = PPI, $R(A, C)$ = PPI, and $R(B, C)$ = PPI, and the center of outer sphere B was assumed to fall into 1NE or 2NE. This situation does not exist in the real world. Under Constraints 2 and 4, there is no solution to this situation. However, the $R5_3$DOS-intersection model can explain the situation that cannot be realized in the 3D space.

Next, the $R5_3$DOS-intersection model was found capable of expressing situation that cannot be illustrated by the MBR model through the analysis of the following example. For any three external spheres A–C in the 3D space, it is assumed that the topological and azimuth relationships between them are known, and these spheres are separated from each other.

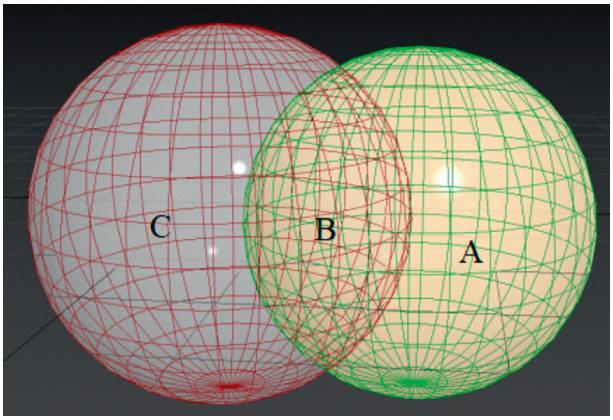


Figure 10: Special case.

For the MBR model, Example 1: (a) dir (A, B)=(1, 1, 1) and (b) dir (A, C)=(1, 1, 1) were obtained for the two examples (Figures 11 and 12).

For the R53DOS-intersection model, layer R5 can be described as $R(A, B) = \mathrm{DR}$, $R(A, C) = \mathrm{DR}$, and $R(B, C) = \mathrm{DR}$. Then,

$$R(A,B) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap C^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon(A^o \cap B^o \cap (C^c)^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix},$$

$$R(A,B) = \begin{pmatrix} \varepsilon(A^o \cap B^o) & (\varepsilon A^o \cap (B^c)^o) \\ \varepsilon((A^o)^o \cap B^o) & \varepsilon((A^o)^C \cap (B^c)^o) \end{pmatrix} = (0 \;\; 1 \;\; 1 \;\; 1),$$

$$R(A,C) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon((A^C)^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix},$$

$$R(A,C) = \begin{pmatrix} \varepsilon(A^o \cap C^o) & (\varepsilon A^o \cap (C^c)^o) \\ (\varepsilon(A^c)^o \cap C^o) & (\varepsilon(A^c)^C \cap (C^c)^o) \end{pmatrix} = (0 \;\; 1 \;\; 1 \;\; 1),$$

$$R(B,C) = \begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) \\ \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon((A^C)^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix},$$

$$R(B,C) = \begin{pmatrix} \varepsilon(B^o \cap C^o) & \varepsilon(B^o \cap (C^c)^o) \\ \varepsilon((B^c)^o \cap C^o) & \varepsilon((B^o)^C \cap (C^c)^o) \end{pmatrix} = (0 \;\; 1 \;\; 1 \;\; 1). \tag{22}$$

Without changing the positions of A–C, the images of the R5$_3$DOS-intersection model in the two examples can be obtained as Figures 13 and 14, where green, blue, and red balls are the outer spheres A–C, respectively.

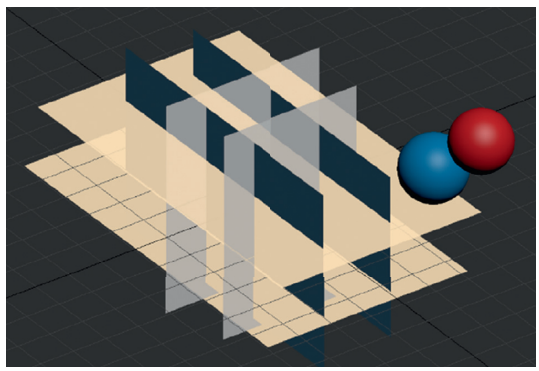


Figure 11: MBR model in Example 1(a).

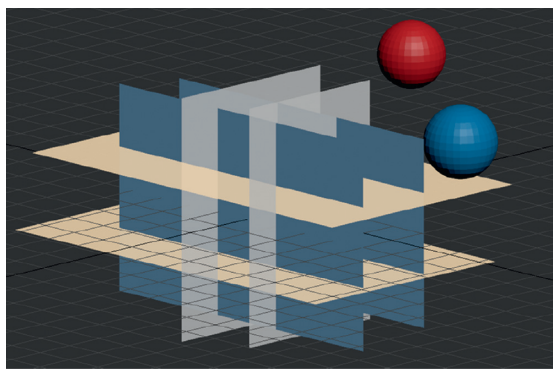


Figure 12: MBR model in Example 1(b).

$$R5_3\mathrm{DOS}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

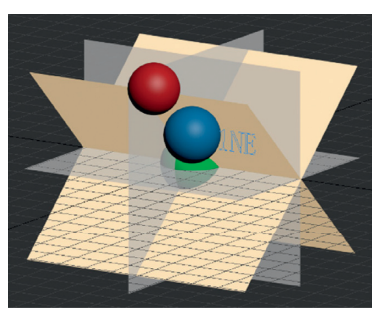


Figure 13: $R5_3$DOS-intersection model in Example 1(a).

$$R5_3\mathrm{DOS}_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

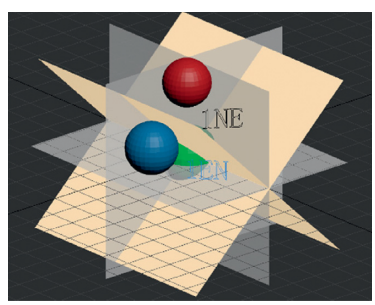


Figure 14: $R5_3$DOS-intersection model in Example 1(b).

$$R5_3\mathrm{DOS}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{23}$$

$$R5_3\mathrm{DOS}_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{24}$$

Example 2. (a) dir (A, B)=(0, 0, 1) and (b) dir $(A, C) = (0, 1, 1)$ were obtained for the two examples (Figures 15 and 16).

In the same way, we can get the corresponding $R_5$3DOS-intersection model (Figures 17 and 18):

$$R5_3\mathrm{DOS}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{25}$$

$$R5_3\mathrm{DOS}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{26}$$

Through the above comparison, it can be seen that the $R_5$3DOS-intersection model can represent the topological relationship of space objects A, B, and C, and it can accurately represent the spatial situation that the MBR model cannot represent.

3.2. Comnd Relationship Reasoning Based on $R5_3$DOS-Intersection Model. This section applies the $R5_3$DOS-Intersection Model to the reasoning of the compound relationships between simple regions in the 3D space. It is assumed that the topological and azimuth relationships between simple regions A and B and those between simple regions B and C are known in advance. Then, the goal is to deduce the possible topological and azimuth relationships between simple regions A and C.

According to Section 2.3, we have

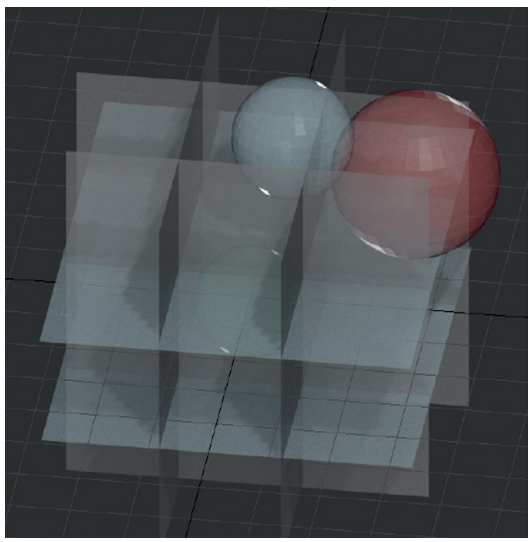


Figure 15: MBR model in Example 2(a).

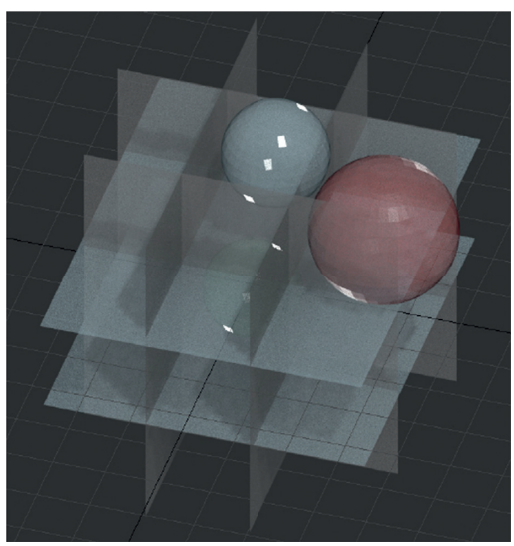


Figure 16: MBR model in Example 2(b).

$$R5_3\mathrm{DOS}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

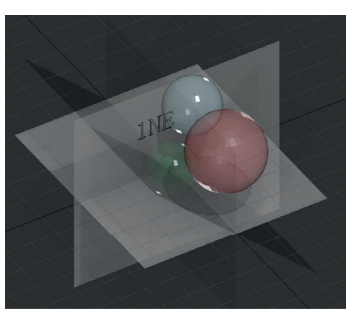


Figure 17: $R5_3$DOS-intersection model in Example 2(a).

$$R5_3\mathrm{DOS}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

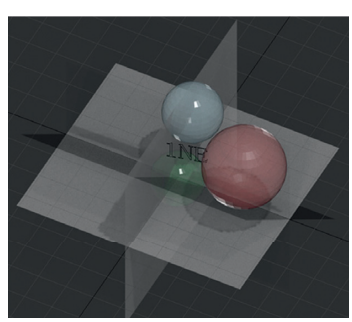


Figure 18: $R5_3$DOS-intersection model in Example 2(a).

Table 3: The list of all directional and topological relationships.

Topological relationships	Regions
DR	DRs1NE, DRs2NE, DRs1EN, DRs2EN, DRs5ES, DRs6ES, DRs5SE, DRs6SE, DRs8SW, DRs7SW, DRs8WS, DRs7WS, DRs4WN, DRs3WN, DRs4NW, DRs3NW
PO	POs1NE, POs2NE, POs1EN, POs2EN, POs5ES, POs6ES, POs5SE, POs6SE, POs8SW, POs7SW, POs8WS, POs7WS, POs4WN, POs3WN, POs4NW, POs3NW
PP	PPs1NE, PPs2NE, PPs1EN, PPs2EN, PPs5ES, PPs6ES, PPs5SE, PPs6SE, PPs8SW, PPs7SW, PPs8WS, PPs7WS, PPs4WN, PPs3WN, PPs4NW, PPs3NW
EQ	Equal
PPI	PPIs1NE, PPIs2NE, PPIs1EN, PPIs2EN, PPIs5ES, PPIs6ES, PPIs5SE, PPIs6SE, PPIs8SW, PPIs7SW, PPIs8WS, PPIs7WS, PPIs4WN, PPIs3WN, PPIs4NW, PPIs3NW

Table 4: Compound relationship reasoning table.

$R(B, C)$	$R(A, B)$	$R(A, C)$
DR (B, C)	all (A, B)	Φ
	POall (A, B)	DRall (A, C); POall (A, C); PPIall (A, C)
	PPall (A, B)	all (A, C)
	EQall (A, B)	all (A, C)
	PPIall (A, B)	DRall (A, C); POall (A, C); PPIall (A, C)
PO (B, C)	DRall (A, B)	DRall (A, C); POall (A, C); PPall (A, C)
	POall (A, B)	Φ
	PPall (A, B)	DRall (A, C); POall (A, C); PPall (A, C)
	EQall (A, B)	POall (A, C)
	PPIall (A, B)	POall (A, C); PPIall (A, C)
PP (B, C)	DRall (A, B)	DRall (A, C); POall (A, C); PPall (A, C)
	POall (A, B)	POall (A, C); PPall (A, C)
	PPall (A, B)	PPall (A, C)
	EQall (A, B)	PPall (A, C)
	PPIall (A, B)	POall (A, C); EQall (A, C); PPall (A, C); PPIall (A, C)
EQ (B, C)	DRall (A, B)	all (A, C)
	POall (A, B)	POall (A, C)
	PPall (A, B)	PPall (A, C)
	EQall (A, B)	EQall (A, C)
	PPIall (A, B)	PPIall (A, C)
PPI (B, C)	DRall (A, B)	DRall (A, C)
	POall (A, B)	DRall (A, C); POall (A, C); PPIall (A, C)
	PPall (A, B)	Φ
	EQall (A, B)	PPIall (A, C)
	PPIall (A, B)	PPIall (A, C)

$$R5_2\text{DOS} = \begin{pmatrix} A^o \cap B^o & A^o \cap (B^c)^o & (A^c)^o \cap B^o & (A^c)^o \cap (B^c)^o \\ s1\text{NE} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s1\text{EN} & s2\text{EN} & s3\text{NW} & s4\text{NW} \\ s5\text{ES} & s6\text{ES} & s7\text{WS} & s8\text{WS} \\ s5\text{SE} & s6\text{SE} & s7\text{SW} & s8\text{SW} \end{pmatrix}. \tag{27}$$

Using the $R5_3$DOS-intersection model, a total of 65 topological and azimuth relationships were obtained from the real world. Hence, it is possible to obtain 65 0-1 matrices of 5 rows and 4 columns, which is denoted as $\Omega_1 = \{R_i; i = 1, \ldots, 65\}R_i$. Targeting at region A, the topological and directional relationships between A and C and those between B and C were taken into account.

Since the topological and azimuth relationships between simple regions A and B and those between simple regions B and C are known in advance, we have $R(B, C) \in \Omega_1$. Then, the possible topological and orientation relationships between A and C were derived from the $R5_3$DOS-intersection model. According to Definition 2, we have

$$
\begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap C^o) \\ \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon(A^o \cap B^o \cap (C^c)^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) & \varepsilon\left((A^C)^o \cap (B^c)^o \cap (C^c)^o\right) \end{pmatrix},
$$

$$
R(A,B) = \begin{pmatrix} \varepsilon(A^o \cap B^o) & \varepsilon(A^o \cap (B^c)^o) \\ \varepsilon((A^c)^o \cap B^o) & \varepsilon\left((A^o)^C \cap (B^c)^o\right) \end{pmatrix},
$$

$$
\begin{pmatrix} \varepsilon(A^o \cap B^o \cap C^o) & \varepsilon(A^o \cap B^o \cap (C^c)^o) \\ \varepsilon(A^o \cap (B^c)^o \cap C^o) & \varepsilon(A^o \cap (B^c)^o \cap (C^c)^o) \end{pmatrix} \vee \begin{pmatrix} \varepsilon((A^c)^o \cap B^o \cap C^o) & \varepsilon((A^c)^o \cap B^o \cap (C^c)^o) \\ \varepsilon((A^c)^o \cap (B^c)^o \cap C^o) & \varepsilon\left((A^C)^o \cap (B^c)^o \cap (C^c)^o\right) \end{pmatrix}, \tag{28}
$$

$$
R(B,C) = \begin{pmatrix} \varepsilon(B^o \cap C^o) & \varepsilon(B^o \cap (C^c)^o) \\ \varepsilon((B^c)^o \cap C^o) & \varepsilon\left((B^o)^C \cap (C^c)^o\right) \end{pmatrix}.
$$

Suppose the real-world 0-1 matrices satisfy

$$M = \{M_i = 1, \ldots, n\}, \tag{29}$$

Then, all 0-1 matrices must meet:

$$m_i = R(A,B) \vee R(B,C), \quad i = 1, \ldots, n. \tag{30}$$

Hence, the matrix that does not satisfy the condition belongs to the empty set, namely, $M \in \varnothing$. This shows the topological and directional relationships R (A, B) and R (A, C) cannot be compounded. Then, all 0-1 matrices represented in the $R5_3DOS$-intersection model were judged one by one. The duplicates in the set $\{M_i = 1, \ldots, n\}$ were removed, leaving the possible topological and azimuth relationships between A and C.

In theory, there are a total of $65 \times 65 = 4{,}225$ topological-azimuth relationships R (A, B) and R (B, C). On this basis, the compound relationship reasoning table was set up (Table 4).

4. Conclusions

This paper extends the compound directional and topological relationships on the 2D plane to the 3D space and then creates the R53DOS-intersection model. Based on the model, a total of 11,038 directional and topological relationships were calculated. Compared with the MBR model, the proposed model can describe the relationships between simple regions accurately and express the relationships with sufficient clarity. To further improve the model, the future research will consider the impact of simple area boundaries on the model and apply the R5DOS model to the formation control of UAV formations.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant nos. 41601454 and 41671397), Science and Technology Development Project of Jilin Province China (Grant no. 20191001008XH), Science Foundation of Jilin Provincial Education Department China (Grant nos. JJKH20200329KJ and JJKH20190922KJ), Development and Reform Commission Project of Jilin Province China (Grant no. 2020C037-7), and Ecology and Environment Department Project of Jilin Province China (Grant no. 2019-02).

Supplementary Materials

This code is the screening algorithm of the R5DOS-intersection model. The purpose is to screen several matrices theoretically in the model according to the constraints and finally get the algorithm of the matrix that meets the requirements, the result of running the code needs simple processing, not the result of the article. The code is developed based on MATLAB software. (*Supplementary Materials*)

References

[1] L. Wang, L. Zhao, G. Huo et al., "Visual semantic navigation based on deep learning for indoor mobile robots," *Complexity*, vol. 2018, Article ID 1627185, 12 pages, 2018.

[2] Y. Fan, K. Xing, X. Jiang et al., "Fuzzy adaptation algorithms' control for robot manipulators with uncertainty modelling errors," *Complexity*, vol. 2018, Article ID 5468090, 8 pages, 2018.

[3] J. Li, K. Deng, X. Huang, and J. Xu, "Analysis and applications of location-aware big complex network data," *Complexity*, vol. 2019, Article ID 3410262, 2 pages, 2019.

[4] D. A. Randell and T. Cohn, *Modelling Topological and Metrical Properties*, University of Warwick, Coventry, UK, 1989.

[5] D. A. Randell, Z. Cui, and A. G. Cohn, "A spatial logic based on regions and connection," in *Proceedings of the 3rd International Conference on Principles of Knowledge Representation*, pp. 165–176, Cambridge, MA, USA, 1992.

[6] M. J. Egenhofer and R. D. Franzosa, "On the equivalence of topological relations," *International Journal of Geographical Information Systems*, vol. 9, no. 2, pp. 133–152, 1995.

[7] M. J. Egenhofer and R. D. Franzosa, "Point-set topological spatial relations," *International Journal of Geographical Information Systems*, vol. 5, no. 2, pp. 161–174, 1991.
[8] S. Li, *Research on Orientation Relations and Integrative Reasoning with Topological Relations and Orientation Relations in Dynamic Settings*, Jilin University, Changchun, China, 2013.
[9] Z. Chen, X. Liu, J. Yang, E. Little, and Y. Zhou, "Deep learning-based method for sem image segmentation in mineral characterization, an example from duvernay shale samples in western Canada Sedimentary Basin," *Computers & Geosciences*, vol. 138, Article ID 104450, 2020.
[10] X. Li, C. Lin, and X. Xu, "A target tracking model for enterprise production monitoring system based on spatial information and appearance model," *Traitement du Signal*, vol. 36, no. 4, pp. 369–375, 2019.
[11] F. Lei, J. Cai, Q. Dai et al., "Deep learning based proactive caching for effective wsn-enabled vision applications," *Complexity*, vol. 2019, Article ID 5498606, 12 pages, 2019.
[12] B. Beirami and M. Mokhtarzade, "Spatial-spectral random patches network for classification of hyperspectral images," *Traitement du Signal*, vol. 36, no. 5, pp. 399–406, 2019.
[13] M. Z. Huang, D. Tian, and H. B. Liu, "A hybrid fuzzy wavelet neural network model with self-adapted fuzzy -means clustering and genetic algorithm for water quality prediction in rivers," *Complexity*, vol. 2018, Article ID 8241342, 11 pages, 2018.
[14] Z. Li, C. Xu, and C. Liu, "Frequent subtree mining algorithm for ribonucleic acid topological pattern," *Revue d'Intelligence Artificielle*, vol. 33, no. 1, pp. 75–80, 2019.
[15] Q. Liu, X. He, F. Guan et al., "Method and implementation of improving the pointing accuracy of an optical remote sensor using a star sensor," *Traitement du Signal*, vol. 36, no. 4, pp. 311–317, 2019.
[16] Z. H. Wang, H. W. Yan, and Y. C. Yang, "Compound spatial query based on direction and distance relation," *Engineering of Surveying & Mapping*, vol. 23, no. 11, pp. 7–12, 2014.
[17] Y. Liu, X. Gong, and D. Kong, "Spatial reasoning based on 3D-ICSRM model," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2892545, 9 pages, 2019.
[18] R. Hou, T. Wu, and J. J. Yang, "Reasoning with cardinal directions in 3D space based on block algebra," in *Proceedings of the International Conference on Electronic Information Technology and Intellectualization*, pp. 500–511, Guangzhou, China, June 2016.
[19] S. Wang, R. Dong, W. Song, and C. Wang, "Qualitative spatial reasoning with oriented point relation in 3D space," *Chinese Journal of Electronics*, vol. 28, no. 2, pp. 325–330, 2019.
[20] M. A. Cobb, F. E. Petry, and K. B. Shaw, "Fuzzy spatial relationship refinements based on minimum bounding rectangle variations," *Fuzzy Sets and Systems*, vol. 113, no. 1, pp. 111–120, 2000.
[21] Y. B. He and J. Bian, "Research on the model of spatial direction relation in spatial reasoning," *Computer & Digital Engineering*, vol. 38, no. 4, pp. 62–65, 2010.
[22] S.-S. Yu, S.-W. Chu, C.-M. Wang, Y.-K. Chan, and T.-C. Chang, "Two improved k-means algorithms," *Applied Soft Computing*, vol. 68, pp. 747–755, 2018.
[23] C. L. Sabharwal and J. L. Leopold, "Cardinal direction relations in qualitative spatial reasoning," *International Journal of Computer Science and Information Technology*, vol. 6, no. 1, pp. 1–13, 2014.

Hindawi
Complexity
Volume 2020, Article ID 7652496, 9 pages
https://doi.org/10.1155/2020/7652496

Research Article

A Big Data Analytics Approach for Dynamic Feedback Warning for Complex Systems

Wenrui Li,[1,2] Menggang Li,[2,3,4] Yiduo Mei,[5] Ting Li,[1] and Fang Wang[1,2]

[1]*School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China*
[2]*Beijing Laboratory of National Economic Security Early-Warning Engineering, Beijing Jiaotong University, Beijing 100044, China*
[3]*National Academy of Economic Security, Beijing Jiaotong University, Beijing 100044, China*
[4]*Beijing Center for Industrial Security and Development Research, Beijing Jiaotong University, Beijing 100044, China*
[5]*Postdoctoral Programme of China Centre for Industrial Security Research, Beijing Jiaotong University, Beijing 100044, China*

Correspondence should be addressed to Menggang Li; mgli1@bjtu.edu.cn and Ting Li; 15113169@bjtu.edu.cn

Received 23 June 2020; Revised 14 September 2020; Accepted 21 September 2020; Published 6 October 2020

Academic Editor: Abd E. I.-Baset Hassanien

With the development of science and technology, the application of big data is becoming more and more widespread, and it has gradually expanded to various fields such as economy and commerce. Since the 2008 international financial crisis, the mainstream economics has shown deficiencies to a certain extent. On the one hand, the expressions pursued by mainstream economic theories are too strict, restricting its processing capabilities. On the other hand, the linearization method ignores the diversity, complexity, and variability of changes in the economic system, which may ignore the emergence of some serious crises. Due to the increasing distance between theoretical models and practice, theoretical models cannot guide the practice and sometimes even mislead the latter. In this paper, we propose a method of dynamic feedback early warning based on big data, which uses the LPPL model to fit parameters. Finally, we used this method to analyze the case of the A-share disaster. The research results show that the method makes the early warning coefficients of dynamic and complex systems more scientific and accurate.

1. Introduction

A complex system is an organic entity that interacts with various parts, not a simple addition and mechanical combination of parts [1]. Therefore, people's understanding of locality cannot be pieced together to understand the overall laws of complex systems.

In the face of complex systems, another method commonly used by humans is to decompose complex problems layer by layer, cut them into tiny fragments, analyze the fragments in depth, and infer the overall situation. However, the close connection and interaction between the various parts of the system were also destroyed during the cutting process.

Human society is a typical complex system. The study of this system, especially the economic aspects, is more a partial analysis and decomposition study. Various traditional economic theories are mostly concentrated in a certain field, such as demand, supply, and currency. In addition, due to the complexity of economic systems, many analyses cannot be simply proved by data nor can they be summarized by simple causality because in complex systems, many factors are both causes and results. In this way, the interpretation of economic phenomena by various economic theories is not convincing. For example, it has been more than 80 years since the Great Depression of 1929, and economists still do not have a satisfactory explanation of the root cause of the Great Depression and are still arguing. The financial crisis of 2008 allowed this debate to continue to heat up, with each side holding its own word and giving up, and this debate will continue in the future [2, 3].

Economists are immersed in economic phenomena, data, models, and theories. They seem to be the most authoritative people who study and solve economic problems.

But what would happen if we looked at economic issues from a completely unrelated perspective? What would come to a conclusion if someone who was familiar with cybernetics but not familiar with economics observed the economic crisis? First of all, the noneconomic professional may not only study a certain part nor study the system decomposition. First, he considers the system as a "black box" and observes the entire system by observing input and output features. Second, he may study the internal structure and connections of the system.

From the perspective of system control, a periodic economic crisis is a periodic shock. If a system enters a periodic oscillation without external input disturbance, this system is likely to have positive feedback [4]. The shock comes from the internal structure of the system, not the changes in input and output. It is wrong to emphasize that changes in certain parameters of the system cause the entire system to oscillate, such as insufficient demand, excessive supply, insufficient investment, credit expansion, credit contraction, and so on. In order to study the periodic oscillation of the system, the structure of the system must be studied to find the positive feedback mechanism inside the system.

2. Related Work

Risks arise from imbalances in supply and demand, both purely quantitative and quality-effective. The traditional risk determination model [5, 6] is based on the assumptions of rational expectations, price equilibrium, and market competition, but these factors cause the theory to be inconsistent with the real economy. The economic system is a complex adaptive system; for example, in Keynes's demand theory, residents must have complex self-adaptation in order to adjust their decisions, including demand and investment, according to the changing environment. Similarly, in Schumpeter's innovation theory, only if a company has complex self-adaptability can it discover and create nonequilibrium and continuously pursue technological innovation and capital accumulation in pursuit of high profits.

An important method for studying complex adaptive systems in cybernetics is to use a two-way feedback mechanism. Therefore, we propose to use a two-way feedback mechanism in cybernetics to warn of risks in economic systems. We propose that the risk originates from the migration of the internal mechanism of the system (from negative feedback to positive feedback). The construction of dynamic feedback models and early risk warning simulation based on big data platforms is the focus of this study. As shown in Figure 1, we divide the system operation cycle into nonrisk cycle and risk cycle according to the dynamic feedback mechanism [7, 8]. We take a semaphore-based dynamic feedback model as an example; when the model is negative feedback, the dynamic feedback model we propose is the traditional equilibrium model. At this time, the system is in a nonrisk cycle. When the model is positive feedback, it is a self-organizing collaborative model that shows strong reflexive characteristics, at which time the system is in a risk cycle.

Our research shows that the negative feedback system conforms to the traditional model of supply and demand and the positive feedback system conforms to the reflexivity theory proposed by Soros [9, 10].

In addition, the financial bubble [11, 12] refers to an economic phenomenon in which the price of a financial asset (or a series of financial assets) undergoes a wave of rise and the market price reflects greater than the upward movement of its actual value. The phenomenon that prices follow a power law and grow faster than exponential growth is called a bubble. Bubble price momentum has gradually increased over time. According to the direction and intensity of price momentum, we divide bubbles into four categories:

Bubble: the price trend rises, and the price momentum gradually increases, as shown in Figure 2

Negative bubble: the price trend declines, and the price momentum gradually weakens, as shown in Figure 3

Reversing bubble: the price trend drops, and the price momentum gradually increases, as shown in Figure 4

Reverse negative bubble: the price trend rises, and the price momentum gradually weakens, as shown in Figure 5

After the bubble, the self-organization of participants in the market through the consensus of the market will form a positive feedback effect and enter the bubble mode. The bubble is not only self-sufficient but also continuously enlarged (the negative bubble is continuously shrinking, that is, the process of debubbling). When the financial bubble trend reached a critical level, although the main participants agreed, the market could no longer withstand the perturbation of a small number of participants, leading to a collapse. In fact, participating in a bubble and trading in a trend is a rational behavior because the risk of a crash will be compensated by the positive returns brought by the financial bubble. At the same time, once a bubble is formed, it will last for a long time, and the burst of the bubble is just a point at the end of this cycle.

In summary, in recent years, with the emergence of complex systems science, the impact of risk interaction behavior has been gradually discovered. However, the main reason for the difficulty of accurate prediction and early warning of economic risks is that traditional risk prediction relies mainly on traditional mathematical modeling tools and lacks the support of effective theoretical models and data algorithms for the interaction and evolution of complex risks. The development and widespread application of big data technology has made various economic systems accumulate huge amounts of data resources. It also makes multidomain and cross-modal data collection and fusion analysis possible [13]. This provides valuable data resources and technical support for comprehensive perception of complex systems, fusion analysis of risk elements, accurate prediction of risk evolution, and timely warning of risks [14]. At the same time, how to effectively extract the characteristics of risk emergence, mutation, evolution, and outbreak from these economic big data so as to realize the intelligent

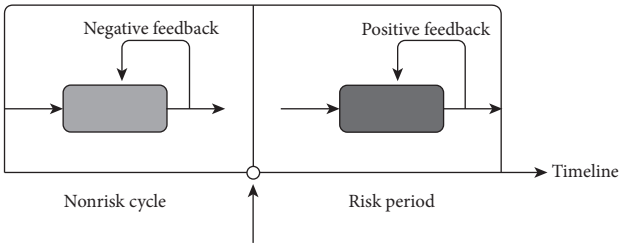


Figure 1: The full risk cycle formed by the feedback mechanism.

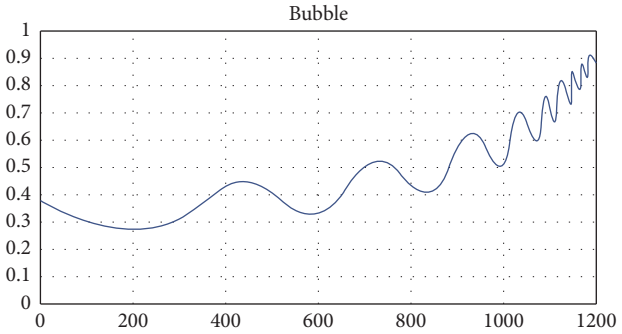


Figure 2: Bubble.

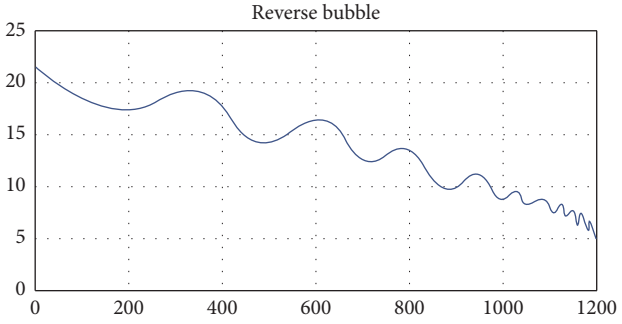


Figure 3: Reverse bubble.

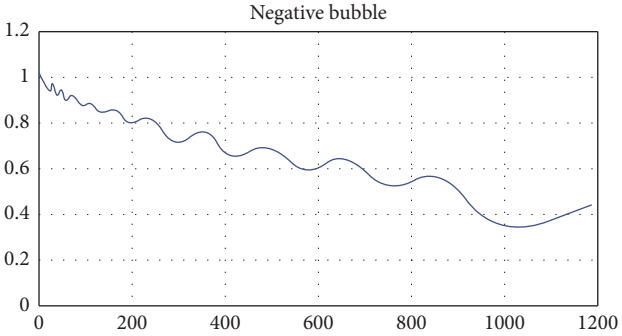


Figure 4: Negative bubble.

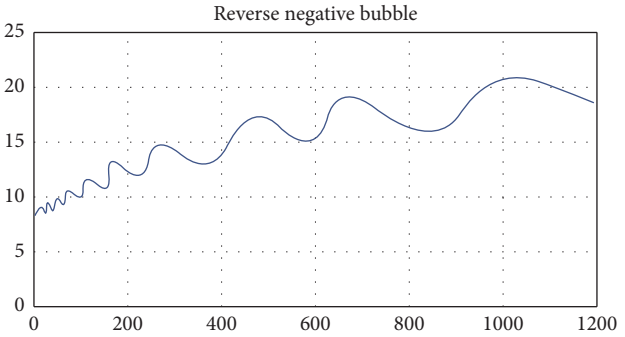


Figure 5: Reverse negative bubble.

prediction and early warning of complex economic risks has increasingly become a new technical challenge.

3. Logarithmic Periodic Power Law Model

Didier Sornette is engaged in the research of geophysics [15]. However, he found that the formation and bursting of financial market bubbles have many similarities with earthquakes and are self-organizing behaviors of complex systems. Professor Sornette then proposed to use the LPPL (log-periodic power law) model (logarithmic periodic power law model) commonly used in the study of geophysics and critical phenomena to study the bubbles in the financial field.

3.1. Related Research. In 1996, Sornette proved through empirical studies that the bubble showed a log-periodic oscillation during the trend towards collapse. In 2001, he further proposed that the stock market bubble can be found in the logarithmic periodic power law model and confirmed this conclusion through empirical research on emerging stock market bubbles. The concept of antibubbles was first proposed by a paper published by Johansen and Sornette in 1999. In this paper, they used a third-order Landau model to predict the trend of the Nikkei [16] in 1999 and successfully predicted it. This reflects the rebound of the Nikkei and its decline again in early 2000, and their predicted rebound is in line with the actual range. Later, Wei-Xing Zhou and Didier Sornette studied the trend of the S & P index after the burst of the new economic bubble at the end of 2002 and found that the U.S. stock market has entered an antibubble stage and further found that there is a global antibubble mechanism in developed western countries. Domestically, Zhang Wei and Huang Xing conducted an empirical analysis of the Shanghai and Shenzhen stock markets through R/S [17], revealing the nonlinear characteristics of China's stock market volatility. Zhou and Professor Sornette observed the "super index" family bubble characteristic of Chinese housing prices [2]. At that time, they predicted based on data up to 20 years, China house prices will rise until 2008, after which the bubble burst. The log-periodic power law model has two common features: one is the logarithmic periodic oscillation. On a linear scale,

the closer to the critical time, the faster the oscillation frequency. The second is power law growth (bubbles) or decay (antibubbles) or superexponential growth or decay, that is, the growth rate of prices is not constant but monotonically increases (bubbles) or decreases (antibubbles). So, it can be considered that the LPPL model gives a quantitative method for judging the bubble and antibubble.

3.2. LPPL Model. Due to the mutual imitation of traders and the formation of the herd effect through positive feedback, the price has a nonlinear vibration, which is similar to a logarithmic periodic vibration so that the final collapse of the bubble risk is explained by market dynamics.

Log periodicity is a significant signal law for discrete scale invariance (DSI) of complex systems. Theoretically, the law of scale invariance often appears near the critical point of a system. Although this phenomenon is not a necessary characteristic for the critical point of the system, the behavior of the critical point of the system can still be observed through the analysis of the law of invariance of the scale. For a variable $O(x)$, its discrete scale invariance characteristic is reflected when the independent variable x becomes the original double. The form of $O(x)$ remains the same, which is $O(x) = \mu(\lambda)O(\lambda x)$. The invariance actually comes from the power law form of $O(x)$. This is consistent with the laws of fractals and multifractals. In order to reflect the periodicity of λ, $O(x)$ can be of the form:

$$O(x) = x^{\alpha}, \tag{1}$$

where $P(\ldots)$ is any function with periodicity. Therefore, this periodic function can be expanded with a Fourier series; the series form is as follows:

$$\sum_{n=-\infty}^{\infty} c_n \exp\left(2n\pi i \frac{\ln x}{\ln \lambda}\right), \qquad a_n = a + i\frac{2\pi n}{\ln \lambda}. \tag{2}$$

Substituting the series expansion into the $O(x)$ formula, the first-order form is

$$I(t) = A + B\tau^{\alpha} + C\tau^{\alpha}\cos[\omega \ln \tau + \phi], \tag{3}$$

where $I(t)$ is the price time series of the financial market, which can be the original price series or the logarithmic price series, $I(t) = P(t)$ or $I(t) = \ln P(t)$; the final fitting results obtained by the two are almost the same, but generally the logarithmic price series is used. τ is the time interval from the critical point, $\tau = t_c - t$ is a bubble curve, $\tau = t - t_c$ is a reverse bubble curve, and t_c is the tipping point. In order to diverge the risk of a bubble burst at a tipping point and to limit prices, $0 < \alpha < 1$. Sometimes in order to fit more finely, a second-order term is added:

$$I(t) = A + B\tau^{\alpha} + C\tau^{\alpha}\cos[\omega \ln \tau + \phi] + D\tau^{\alpha}\cos[2\omega \ln \tau + \phi_2], \tag{4}$$

where $A = \ln[\omega(t)] > 0$ means that if the bubble continues to a critical time t_c, the value of $\ln P(t_c)$ will be reached; B refers to the increase in $\ln P(t_c)$ per unit time when C is close to 0 before the crash; C is a measure of the fluctuation amplitude around the exponential growth; $t_c > 0$ is the time when the bubble bursts and the critical time $t > t_c$ is any time before the bubble bursts; m is the exponential of power growth; ω is the angular frequency of the bubble phase oscillation; and $0 < \phi < 2\pi$ is the initial phase of the shock.

The logarithmic periodic power law model is a relatively successful model developed by economic physicists from the observation of the economic bubble phenomenon [18, 19] and has been adopted by many financial institutions. It is inherently difficult to analyze and make predictions about the bubble trend in financial markets. Economists often make predictions just when the bubble is about to burst. Sornette made an analogy from the earthquake research and analyzed the formation of bubbles in the financial market in depth. He proposed the LPPL model and was able to predict the critical time point. However, as a complex system, financial markets have scale invariability near the critical point, but in the actual development process, how to use models to avoid risks and minimize the losses caused by the burst of the bubble still requires many people to seriously consider things.

3.3. Fitting Process. Essentially,

$$y_i = A + B(t_c - t_i)^{\beta} + C(t_c - t_i)^{\beta}\cos[\omega \ln(t_c - t_i) + \phi]. \tag{5}$$

This is a highly nonlinear equation with 7 parameters to be evaluated. We generally use curve fitting to get the parameters of the model. Faced with as many as 7 parameters to be estimated in the model, we need to fit these parameters, and highly nonlinear functions need to be carefully considered, and what is considered is to ensure that the best fitting results are obtained (underprepared fit or overfit is not good).

We know that the so-called parameter fitting [20] is nothing more than an objective function minimization problem. Here, our objective function is the sum of the squares of the errors between the calculated value Y_i of the LPPL model and the actual observed value (price):

$$\mathrm{SE} = \sum_{t=t_1}^{t_n} (y_t - \hat{y}_t)^2 = \sum_{t=t_1}^{t_n} \left\{y_t - A - B(t_c - t)^{\beta}\left(1 + C\cos(\omega \log(t_c - t) + \varnothing)\right)\right\}^2. \tag{6}$$

When there are too many parameters to be estimated, the objective function will have multiple local minima. In addition, because of the noisiness of the fitting sample, we directly fit the 7 parameters, and we will fall into the misunderstanding of local solution. In previous studies, many

scholars suggested using 1stopt [21, 22] in China to perform 7-parameter fitting. However, research shows that even a general global algorithm such as 1stopt, which does not need to set an initial value, cannot guarantee that the result obtained is the global minimum. After many experiments, one of the alternative ideas is to use the variant 1stop solution:

$$(y_i - A)(t_c - y_i)^{-\beta} = B + C\cos[\omega \ln(t_c - t_i) + \varnothing]. \quad (7)$$

In fact, we used the default LM + UGO algorithm to perform 10 fittings, and the parameters obtained obviously converge to 3 different intervals because if there is no range constraint, there will be multiple sets of solutions in theory.

Therefore, trying to reduce the number of parameters to be estimated becomes the first task of LPPL fitting. To reduce the number of fitted free parameters, the three linear parameters (A, B, and C) in the LPPL are compressed to the remaining four nonlinear parameters (t_c, β, ϕ, and σ). According to the objective function, after the partial derivatives of three linear parameters (A, B, and C) are obtained, the derived derivative formula should be 0 when the minimum value is obtained, and we can get simultaneous equations.

Specifically, first rewrite the LPPL equation as follows:

$$\begin{aligned} y_i &= A + Bf_i + Cg_i, \\ f_i &= (t_c - t_i)^{\beta}, \\ g_i &= (t_c - t_i)^{\beta}\cos[\omega \ln(t_c - t_i) + \phi]. \end{aligned} \quad (8)$$

Then, the equations can be obtained as follows:

$$\begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i f_i \\ \sum_{i=1}^{N} y_i g_i \end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^{N} f_i & \sum_{i=1}^{N} g_i \\ \sum_{i=1}^{N} f_i & \sum_{i=1}^{N} f_i^2 & \sum_{i=1}^{N} f_i g_i \\ \sum_{i=1}^{N} g_i & \sum_{i=1}^{N} f_i g_i & \sum_{i=1}^{N} g_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}. \quad (9)$$

The above equations can be solved by standard LU decomposition [23] to obtain A, B, and C. In this way, we can use 4 nonlinear parameters to represent 3 linear parameters. After the operation of the aforementioned slave parameters, we reduced the number of parameters to be estimated from 7 to 4.

After that, we generally use the taboo + LM algorithm to perform the fitting to get the final result.

4. Case Study of A-Share Disaster

China's A-share market has ushered in a long-lost bull market after a 7-year bear market. Due to the typical leverage characteristics of this round of bull market, in less than a year, the transaction volume of the two markets has exceeded 1 trillion to 20,000. The index also raised a double within one year. Fanatic funds have blown up a huge bubble in the A-share market. The stock market bubble is often caused by the herd effect of investors [24, 25]. Therefore, we can observe China's A-share market by describing indicators related to investor behavior in the stock market to see the bubble situation of the stock market.

4.1. Bubble Burst. The Chinese economy has shifted from being largely closed to being a major player in the world, mainly through large-scale investment in real estate and infrastructure over the past three decades. From 2000 to 2008, China's average annual GDP growth rate was 13%, but it has slowed to 7.8% in 2009 and around 7% after 2015. With this rapid growth, from May 2005 to October 2007 and from November 2008 to August 2009, the Chinese stock market experienced dynamic changes in the roller coaster. The latest bubble started around mid-2014 and recently began to collapse in mid-June 2015, commonly known as the stock market disaster.

The last bubble reached almost 150% growth in just one year. As China's real estate market cools down sharply as a whole, China's stock market has grown even more dramatically. This bubble can be seen as the result of powerful leverage that is disconnected from the reality of economic activity and corporate returns. About 7% of China's population has been active in the stock market frenzy, from easy access to credit to investing in the stock market.

An interesting feature of the Chinese stock market is that insurance companies and pension funds have traditionally stabilized investors through buying and holding strategies, which basically do not exist in the field of Chinese investment. As a result, about 90 million small- and medium-sized investors have become the main driving force of the stock market and are more vulnerable to rumors, imitations, speculation, and crowds. In fact, there are many signs that the Chinese government has encouraged small retail investors to invest in the stock market, driving it for some time, but also catalyzing its vulnerability.

4.2. Early Warning and Analysis of Bubble Burst. Since the high point on June 12, 2015, the SSEC has fallen by 32% since reaching the bottom on July 8, 2015, and has fluctuated significantly since then. The smaller Shenzhen stock market has fallen 41% over the same period. The Chinese government has taken unprecedented measures to stop the decline. In particular, the People's Bank of China has repeatedly lowered the benchmark loan and deposit interest which was at an all-time low point. In addition, more measures have been taken like relaxing funding rules, announcing a moratorium on suspension of new IPO, investigating malicious short selling and so on. Before the stock index began to fall sharply in mid-June, there were obvious signs in the A-share market, especially the daily volatility, turnover rate, and unprecedented trading volume.

Taking the last trading day in May as an example, the amplitude throughout the day is as high as 300 points, which is more than 7%. Then, entering the middle of June, the market began a plunge pattern. SSE Composite Index fell from a maximum of 5178 to 3507, a drop of 31%. The GEM index and the SME index were also spared. The SME index fell 33%, and the GEM index plummeted 40%. During the stock market disaster, the total market value of A shares lost

a total of 25 trillion yuan, so that the surge in A shares from early 2015 to mid-June was brought to naught.

When the fuse was lit after the SEC cracked down on OTC allocations, the market entered a plunge mode. First, a large number of funding accounts were forcibly closed. Because of the leverage of 1 : 5 or even 1 : 10, a slight breeze in the market is also unacceptable for such accounts. When a large number of funded accounts reach the warning line, in order for the funded company to protect its own funds, the computer will automatically sell a large number of stocks, leading to the emergence of a limit tide. Under the slump, the Chinese government has adopted several rounds of rescue measures, including interest rate cuts and transaction costs, but these measures failed because they did not touch the essence of the crisis.

The failed bailout triggered a frontal collapse of confidence and a second round of plunge, which led to the collapse of the two financial accounts of major securities firms and a large number of liquidation lines. A large number of liquidation positions were sold on the daily limit board, which eventually formed a vicious circle. On the other hand, the net values of major private equity companies like Qingshuiyuan, continued to fall during the crash. Some of them reaching the liquidation line 0.7, were forced to liquidate. Larger public offering funds are facing a severe tide of redemption, and the sharp sell-off of assets in an inherently illiquid market is undoubtedly fueling the fire.

As the stock market continued to plummet in mid-June, the Chinese government began to implement a number of market rescue policies in order to stabilize the market and prevent the occurrence of local financial crises [26–28]:

(1) On June 26, the Shanghai Composite Index recorded a 74% drop. The next day, the People's Bank of China announced that starting from June 28, financial institutions would cut interest rates by 0.25%, while targeting downwards by 0.256.

(2) Two days later, due to the market's downward trend, the two major exchanges announced at the same time that they would reduce the transaction costs of the stock market.

(3) After the plunge, the CSRC announced that it would relax the restrictions on margin financing and securities lending. The discount rate and the liquidation line can be determined by the securities firms.

(4) On July 3, the CSRC issued an announcement that it was necessary to reduce the number of new shares issued under the current situation.

(5) On July 3, the China Securities Regulatory Commission issued an announcement saying that Central Huijin Investment Co., Ltd. has invested funds to rescue the market.

(6) After one day, a number of securities companies jointly expressed their confidence in the market and actively used their own funds to support the market.

(7) On July 4, more than 20 large securities firms announced joint funding to purchase not less than 120 billion yuan of financial real estate index constituents in the secondary market.

(8) On the evening of July 4, under the influence of the State Council, 28 companies that had planned to be listed on the Shanghai and Shenzhen stock exchanges were suspended from IPO, and all investors were returned with frozen subscription funds.

(9) On July 5, 2015, the People's Bank of China announced that it will provide adequate liquidity support on the balance sheet of China Securities Finance Corporation, and on the same day, the China Securities Regulatory Commission confirmed that China Securities Finance Corporation would buy ETF.

(10) On July 5, 2015, the China Financial Futures Exchange imposed restrictions on opening futures markets, especially short positions, and imposed penalties on naked short positions.

From the stock disaster to the rescue of the market, in 2015, A shares staged a war that is hard to see in Hollywood. Among those strategies, the sweeping of stocks by 'the State Team' is the most noticeable. The "Equilibrium Fund" swept A shares in the process of saving the market and gave birth to rebounding demonic stocks such as Meibang Clothing and Luoyang Glass. According to Wande data, as of the end of September, among the top ten circulating shareholders of China's A-share companies, China Securities Finance Co., Ltd. and Central Huijin Investment Co., Ltd. appeared in more than 1,300 of them, accounting for about 49% (560 from the Shanghai Stock Exchange, 232 from the Shenzhen Stock Exchange, 353 from the SME Board, and 217 from the GEM). The amount of funds used to purchase A shares accounted for about 8% of the total A shares outstanding.

4.3. Descriptive Statistical Characteristics of A-Share Market Returns. From the statistics of returns of Table 1, it can be seen that the yield index of the Shanghai Composite Index has the characteristics of spikes and fat tails, that is, the probability of extreme returns is large.

4.4. Significance of Liquidity. Market liquidity can be defined from two aspects: one is the liquidity of assets and the other is the liquidity of liabilities. In short, the liquidity of assets is the time cost and money cost of turning it into money. Time cost refers to the time it takes to realize. If investors who urgently need currency need to obtain it after one-month liquidity, then this asset is obviously not a liquid asset. At the same time, the high cost of money for realizing money also means low liquidity, and the liquidity of liabilities is the difficulty and cost of financing. In case of the A-share market, as long as the stock is not at the limit, the time and cost of liquidation can be ignored. However, due to the design of the $T+1$ trading system, the liquidity of the A-share market is naturally lower than the mature $T+0$

Table 1: Descriptive statistics of A-share returns.

Mean	0.097819
Median	0.37415
Maximum value	5.7635
Minimum value	−8.4909
Standard deviation	2.636989
Skewness	−0.792691
Kurtosis	4.196858

market because the stocks bought on that day can only be sold the next day, which means that the overall time cost of A shares is higher than that in other markets.

4.5. Analysis. The occurrence of this round of stock disasters was a series of liquidity crises caused by leveraged funds entering the stock market, financing, and other account bursts [29–31].

Looking back at the policies implemented by the government at the time of the stock market crash, most of them did not touch on the substance of the problem, but instead, the collapse of confidence caused by the ineffectiveness of the policy led to a second decline. For example, the entry of pensions into the market is a long-term process. It is impossible to require pensions to pick up the market for the country, resulting in deeper lock-in and causing greater problems. Therefore, the entry of pensions into the market has no effect on the settlement of short-term liquidity. Reducing transaction settlement costs is even more irrelevant. The suspension of the IPO has eased the thawing of some financing funds, but it has not had a great effect on the problems on the market. More than 20 brokerage companies invested 120 billion yuan to buy blue-chip ETFs and to lift PetroChina by the end of the year, and other measures could not solve the SME board and the ChiNext board in the hard-hit areas of liquidity, as well as various small- and medium-sized market value stocks. In addition, increasing the QFII quota has not shown good results because foreign capital comes in time, and it does not come to play a role in unwinding the financing disk. Even for a small interest rate cut and RRR cut, the effect is not significant because the release of the interest rate cut effect takes time, and the market's expectations of most interest rate cuts have been reflected in the stock price. At this time, interest rate cuts cannot drive the stock market out of liquidity crisis.

The bull market marked by the behavior of risk-free interest rates has experienced a surge in interest rates, and the logic of being long has collapsed. At the same time, leveraged funds have grown and cannot be ignored. Sensitive speculators will lose their way with a bit of wind. For highly leveraged accounts, one or two daily stops will lead to an automatic liquidation. So far, a model of a liquidity spiral has been formed as shown in Figure 6.

In contrast, after the market closed on July 8th, the securities company injected 200 billion yuan into five public funds, immediately welcoming the positive response of the Hong Kong stock market at night. Moreover, the stock

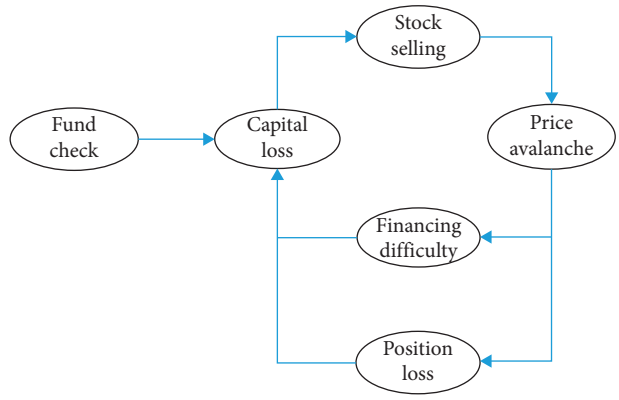


Figure 6: Liquidity spiral.

market rose sharply for two consecutive days since then, and the panic was relieved. This is a result of the policy responds positively to the liquility issues. These 200 billion yuan can help ease the huge redemption tide facing the fund and cut off the spread of the liquidity crisis. Moreover, compared with government officials, public fund practitioners know the market better and know what stocks to buy, which is more reasonable than Huijin's direct purchase of large corporate stocks.

In view of this, in the future, if a liquidity crisis occurs again, this article proposes the following short-term emergency measures:

(1) The People's Bank of China publicly promises to provide sufficient liquidity to major commercial and securities companies.

(2) The People's Bank of China shall lead the establishment of the leveling fund, which must be large in scale and fast in timing.

(3) Prohibit malicious suspension of trading to prevent secondary depletion of liquidity. The above has analyzed the bad consequences caused by the malicious suspension of trading during the 2015 A-share stock disaster.

The entire stock disaster went from brewing to occurrence to deterioration to mitigation. It can be said that various policy errors have occurred. To avoid the recurrence of future tragedies, various systems need to be established and perfected. Therefore, this article proposes the following three policy recommendations for discussion:

(1) It can also be seen in the analysis of the entire stock market disaster that in the beginning of June before the stock market disaster, the market was already in an extremely bubbling state. At this time, the interest rate of the money market rose sharply due to factors such as seasonal bank deposits. The central bank has tightened liquidity through targeted measures, but the CSRC chose to strictly check the funding at this time instead of choosing a time when these funds were not tight in March and April. It can be said that because the supervision system of the three parties

and the three associations also has greater responsibility for this stock disaster, it is urgent to establish a comprehensive financial supervision institution.

(2) Improve the existing arbitrage system [32, 33] to avoid loopholes in the trading system.

(3) Leveraged funds played an important role in this round of the stock disaster. We must learn from domestic and foreign experiences to improve the supervision of leveraged funds. The United States has many years of history of funding, but they have strict restrictions on the subject of funding. Many penny stocks and too high-risk stocks cannot be bought in leveraged accounts [34]. At the same time, the source of leveraged funds is not only securities firms, but also banks, trusts, private loans, etc., so comprehensive financial regulators are required to supervise cross-border funds.

5. Conclusive Remarks

In this paper, we propose a method of dynamic feedback early warning based on big data, which uses the LPPL model to fit parameters. We evaluate the performance of real-time diagnostics, economic theory bubbles based on rational expectations, imitation of behavioral mechanisms, and the mathematical expressions. Finally, we used this method to analyze the case of the A-share disaster. The research results show that the method makes the early warning coefficients of dynamic and complex systems more scientific and accurate to meet the needs of China's financial sector at this stage.

In the future, more big data technologies will be considered for dynamic complex system early warning. For example, a unified open source platform for distributed online machine learning will be built to collect and analyze data in the economic field; more prediction models will be integrated into the platform as an analysis lib tool; more real-time suggestions will be given to use big data analysis and comprehensive decision-making methods. These works will bring new dividends to the early warning of complex systems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Program of the Co-Construction with the Beijing Municipal Commission of Education of China (grant no. B18H100040).

References

[1] M. Mitchell, "Complex systems: network thinking," *Artificial Intelligence*, vol. 170, no. 18, pp. 1194–1212, 2006.

[2] W.-X. Zhou and D. Sornette, "Antibubble and prediction of China's stock market and real-estate," *Physica A: Statistical Mechanics and Its Applications*, vol. 337, no. 1-2, pp. 243–268, 2004.

[3] M. Crouhy, G. Dan, and R. Mark, "A comparative analysis of current credit risk models," *Journal of Banking & Finance*, vol. 24, no. 1-2, pp. 59–117, 2000.

[4] D. Sornette, *Why Stock Markets Crash: Critical Events in Complex Financial Systems*, Vol. 49, Princeton University Press, Princeton, NJ, USA, 2017.

[5] S. Davis, "The adverse feedback loop and the effects of risk in both the real and financial sectors," vol. 66, Federal Reserve Bank of Dallas, Washington, DC, USA, 2010.

[6] I. Logojan, "Explanations of financial crises in the classic theory and in the theory of reflexivity," *Review of the Air Force Academy*, vol. 2, p. 71, 2009.

[7] I. Vrecko, J. Kovac, B. Rupnik, and B. Gajsek, "Using queuing simulation model in production process innovations," *Int. Journal of Simulation Modelling*, vol. 18, no. 1, pp. 47–58, 2019.

[8] D. T. Pele, "An LPPL algorithm for estimating the critical time of a stock market bubble," *Journal of Social and Economic Statistics*, vol. 1, no. 2, pp. 14–22, 2012.

[9] A. Johansen and D. Sornette, "Financial "anti-bubbles": log-periodicity in gold and Nikkei collapses," *International Journal of Modern Physics C*, vol. 10, no. 4, pp. 563–575, 1999.

[10] L. M. Wang, Z. Y. Hao, X. M. Han, and R. H. Zhou, "Gravity theory-based affinity propagation clustering algorithm and its applications," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 4, pp. 1125–1135, 2018.

[11] Z. Ma and S. Xiao, "Closed form valuation of vulnerable European options with stochastic credit spreads," *Economic Computation And Economic Cybernetics Studies And Research*, vol. 53, no. 4, pp. 293–311, 2019.

[12] F. J. Carmone and P. E. Green, "Model misspecification in multiattribute parameter estimation," *Journal of Marketing Research*, vol. 18, no. 1, p. 87, 1981.

[13] D. Blazquez and J. Domenech, "Big data sources and methods for social and economic analyses," *Technological Forecasting and Social Change*, vol. 130, pp. 99–113, 2018.

[14] V. G. Chumak, V. M. Ramzaev, and I. N. Khaimovich, "Challenges of data access in economic research based on big data technology," *CEUR Workshop Proceedings*, vol. 1490, 2015.

[15] N. M. S. Algheriani, V. D. Majstorovic, S. Kirin, and V. Spasojevic Brkic, "Risk model for integrated management system," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1833–1840, 2019.

[16] G. Shabat, Y. Shmueli, A. Averbuch, and Y. Aizenbud, "Randomized lu decomposition," *Applied & Computational Harmonic Analysis*, vol. 44, no. 2, 2013.

[17] E. Comăniţă, P. Cozma, I. Simion, M. Roşca, and M. Gavrilescu, "Evaluation of eco-efficiency by multicriteria decision analysis. case study of eco-innovated and eco-designed products from recyclable waste," *Environmental Engineering and Management Journal*, vol. 17, pp. 1791–1804, 2018.

[18] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation*

and Economic Cybernetics Studies And Research, vol. 53, no. 2, pp. 77–92, 2019.

[19] M. Spatareanu, V. Manole, and A. Kabiri, "Do bank liquidity shocks hamper firms' innovation?" *International Journal of Industrial Organization*, vol. 67, Article ID 102520, 2019.

[20] S. Figlewski, "Hedging performance and basis risk in stock index futures," *The Journal of Finance*, vol. 39, no. 3, pp. 657–669, 1984.

[21] D. Zhang, "High-speed train control system big data analysis based on fuzzy RDF model and uncertain reasoning," *International Journal of Computers, Communications & Control*, vol. 12, no. 4, 2017.

[22] D. Zhang, J. Sui, and Y. Gong, "Large scale software test data generation based on collective constraint and weighted combination method," *Tehniki Vjesnik*, vol. 24, no. 4, pp. 1041–1049, 2017.

[23] S. Jiang, M. Lian, C. Lu, Q. Gu, S. Ruan, and X. Xie, "Ensemble prediction algorithm of anomaly monitoring based on big data analysis platform of open-pit mine slope," *Complexity*, vol. 2018, Article ID 1048756, 13 pages, 2018.

[24] S. Hido, S. Tokui, and S. Oda, "Jubatus: an open source platform for distributed online machine learning," in *Proceedings of the NIPS 2013 Workshop on Big Learning*, Lake Tahoe, NV, USA, December 2013.

[25] B. N. Silva, M. Diyan, and K. Han, "Big data analytics," in *Deep Learning: Convergence to Big Data Analytics*, pp. 13–30, Springer, Singapore, 2019.

[26] B. Jan, H. Farman, M. Khan et al., "Deep learning in big data analytics: a comparative study," *Computers & Electrical Engineering*, vol. 75, pp. 275–287, 2019.

[27] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big data analytics: computational intelligence techniques and application areas," *Technological Forecasting and Social Change*, vol. 153, Article ID 119253, 2020.

[28] M. Haslett, "Dynamic feedback system and method for providing dynamic feedback," U.S. Patent Application 15/6936142019-3-7, 2019.

[29] C. Diks, C. Hommes, and J. Wang, "Critical slowing down as an early warning signal for financial crises?" *Empirical Economics*, vol. 57, no. 4, pp. 1201–1228, 2019.

[30] J. Janekova, J. Fabianova, and M. Fabian, "Assessment of economic efficiency and risk of the project using simulation," *International Journal of Simulation Modelling*, vol. 18, no. 2, pp. 242–253, 2019.

[31] P. Wang, L. Zong, and Y. Ma, "An integrated early warning system for stock market turbulence," *Expert Systems with Applications*, vol. 153, Article ID 113463, 2020.

[32] A. G. Barkish, T. Salari, and N. Salehnia, "Analysis and modeling of the substantial fall of Tehran stock exchange in January 2014 using the log-periodic power law (LPPL) model," *Journal of Economic Research*, vol. 153, pp. 97–124, 2019.

[33] O. I. Krivosheev, "Log-periodic power law autonomous stock market model," in *Proceedings of the 2019 Twelfth International Conference "Management of Large-Scale System Development" (MLSD)*, IEEE, Russia, Moscow, pp. 1–5, October 2019.

[34] S. Harsha and B. Ismail, "Review on financial bubbles," *Statistical Journal of the IAOS*, vol. 35, no. 3, pp. 501–510, 2019.