Wireless Communications and Mobile Computing

Semantic Sensor Data Annotation and Integration on the Internet of Things

Lead Guest Editor: Xingsi Xue
Guest Editors: Yuemin Ding, Pei-Wei Tsai, and Chin-Ling Chen

Semantic Sensor Data Annotation and Integration on the Internet of Things

Wireless Communications and Mobile Computing

Semantic Sensor Data Annotation and Integration on the Internet of Things

Lead Guest Editor: Xingsi Xue
Guest Editors: Yuemin Ding, Pei-Wei Tsai, and Chin-Ling Chen

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6653816, 12 pages
https://doi.org/10.1155/2021/6653816

Research Article

Mining Profitable and Concise Patterns in Large-Scale Internet of Things Environments

Jerry Chun-Wei Lin,[1] **Youcef Djenouri**,[2] **Gautam Srivastava**,[3,4] **and Philippe Fournier-Viger**[5]

[1]*Western Norway University of Applied Sciences, Norway*
[2]*SINTEF Digital, Norway*
[3]*Brandon University, Canada*
[4]*China Medical University, Taiwan*
[5]*Shenzhen University, Shenzhen, China*

Correspondence should be addressed to Jerry Chun-Wei Lin; jerrylin@ieee.org

Received 24 December 2020; Accepted 7 September 2021; Published 23 September 2021

Academic Editor: Xingsi Xue

In recent years, HUIM (or a.k.a. high-utility itemset mining) can be seen as investigated in an extensive manner and studied in many applications especially in basket-market analysis and its relevant applications. Since current basket-market scenario also involves IoT equipment to collect information, i.e., sensor or smart devices, it is necessary to consider the mining of HUIs (or a.k.a. high-utility itemsets) in a large-scale database especially with IoT situations. First, a GA-based MapReduce model is presented in this work known as GMR-Miner for mining closed patterns with high utilization in large-scale databases. The k -means model is initially adopted to group transactions regarding their relevant correlation based on the frequency factor. A genetic algorithm (GA) is utilized in the developed MapReduce framework that can be used to explore the potential and possible candidates in a limited time. Also, the developed 3-tier MapReduce model can be easily deployed in Spark for the handlings of any database of large scale for knowledge discovery of closed patterns with high utilization. We created sets of extensive experimental environments for evaluating the results of the developed GMR-Miner compared to the well-known and state-of-the-art CLS-Miner. We present our in-depth results to show that the developed GMR-Miner outperforms CLS-Miner in many criteria, i.e., memory usage, scalability, and runtime.

1. Introduction

As there is rapid growth of information technologies regarding machine learning models, Internet of Things (IoT) [1], and edge and cloud computing [2, 3], data-driven mining has become an important topic that can be used to extract the meaningful information from the collections of those techniques. Several pattern mining models [4–9] have been extensively studied, and the most fundamental knowledge of pattern mining in knowledge discovery in databases (KDD) is called ARM or association rule mining, which is deployed through varied applications and specific domains. Among them, Apriori was presented for finding the association rules set in transactional databases iteratively. This is a standard approach that finds the candidate itemsets first then derive the satisfied itemsets at each level or called as the level-wise/generate-and-test model; a huge memory usage and the computational cost are relevantly high. After that, a set of association rules can be discovered and mined. Frequent pattern- (FP-) tree [10] was designed to speed up mining progress by building a condense tree structure. Thus, only frequent 1-itemsets are held in the main memory that can be used for later mining progress. In addition, a conditional FP-tree is then recursively constructed to find the frequent itemsets (or frequent patterns) according to different prefix itemsets in the Header_Table. Both Apriori and FP-tree algorithms ensure the DC (or a.k.a. downward closure) property to avoid the heavy cost regarding "combinational explosion."

This property is then applied and extended to many pattern mining algorithms in different domains and applications, i.e., HUIM (or called high-utility itemset mining) [11–15].

HUIM used 2 properties (a.k.a. the internal + external utility) to find the set of HUIs (or a.k.a. high-utility itemsets) in the basket-market domain. The internal utility can be considered as the quantity of an item of each transaction in databases, and external utility can be treated as the unit profit value of each item in databases. Those two values can be replaced by other factors according to the specific requirements, constraints, users' needs, and applications. The generic algorithm of HUIM [16] does not take DC property for revealing the set of HUIs, which requires a huge size of the search space. To solve this limitation, TWU (or a.k.a. transaction-weighted utilization) model [14] considers the transaction utility to construct the HTWUIs (or a.k.a. high transaction-weighted utilization itemsets) as the itemsets with the upper-bound values for maintaining the DC property, which is named as TWDC (or a.k.a. transaction-weighted downward closure) in HUIM. This property is then used in many utility-driven mining algorithms, e.g., UP-growth+ [17], HUI-Miner [15], HUP-tree [18], FHM [19], and d2HUP [20]. More algorithms to improve mining effectiveness regarding the discovered patterns are then developed and discussed by adapting utility concept in pattern mining tasks. In IoT applications [21], many factors can be considered as different values, e.g., interestingness, weight, importance, and uncertainty degree; thus, HUIM can be easily adopted into IoT and/or sensor networks to further discover the required information for data analysis tasks. Based on this assumption, more important and specific information and knowledge will be discovered for later decision or strategy making.

Instead of classic pattern mining approaches such as FIM (or a.k.a. frequent itemset mining) or ARM (or a.k.a. association rule mining) for decision-making, it can disclose more useful and relevant information based on the property of HUIM. The reason is that the HUIM can reveal more information by taking internal and external factors in the mining progress. However, the generic model for discovering the required patterns requires to analyse a huge number of the candidates first, which is inefficient and it is also hard to find the meaningful patterns from a very huge number of patterns. Closed-pattern mining constraint [1, 22–25] was then adapted in pattern mining to provide better functionalities for mining condense and compress patterns. This strategy is then used in HUIM, which is arisen as a new topic called CHUIM (or a.k.a. closed high-utility itemset mining) [26, 27]. Based on this model, less but more meaningful information will be discovered by two conditions as follows: (1) the superset of an itemset has different support values to an itemset itself and (2) the utility of an itemset is no less than the predefined minimum utility count (threshold). CHUD algorithm [26] was investigated to firstly find the CHUIs (or called closed high-utility itemsets) by using the generic TWU model [14]. Since TWU model is a level-wise and generate-and-test model, a huge number of the computational cost is needed and a huge memory usage is required to keep the candidates level-by-level, which is inefficient and time-consuming. CHUI-Miner [27] was investigated to build the extended utility-list (EU-list) that keeps the revealed information in the main memory; the divide-and-conquer mechanism is then used to find the CHUIs correctly and completely. To better improve mining performance, CLS-Miner [28] was designed by using the matrix to lower the size of the search space. This model has good performance compared to the existing models and is considered as the state-of-the-art approach for CHUIM. The generic CLS-Miner is, however, not possible to be performed for discovering the CHUIs in large-scale databases; it is inappropriate in real and industrial domains and applications. Past works have been developed to present the parallel and distributed models used in HUIM [29], but those generic models need to find a very large set of the candidate itemsets for decision-making; it needs high computational cost and a huge memory usage to deliver the complete information. To build an effective and efficient model for revealing the CHUIs has become an important issue in pattern mining research.

Up until now, there has been no model existing that can be used for CHUIM in any database of large scale. Moreover, in the case of correctly and completely mining the needed CHUI making use of distributed and parallel frameworks, we require a strong model to be able to distribute the transactions in an effective and efficient manner to the processing nodes. For solving this known limitation, GMR-Miner is developed and introduced in this paper. Main findings are as follows:

(i) We design a 3-tier MapReduce framework deployed in Spark for mining CHUIs in large-scale datasets

(ii) A k-means model is made use of for grouping relevant transactions into clusters; thus, ensuring discovered CHUI numbers is complete and correct

(iii) A GA-based model makes utilization of the MapReduce framework to explore the possible and potential candidates in a limited time for greatly reducing the computational cost

(iv) Experimental evaluation shows that GMR-Miner has a strong and outstanding performance

2. Related Work

2.1. MapReduce Framework. MapReduce [30] is a parallel and distributed framework that was originally designed and implemented by Dean and Ghemawat. It can be made and implemented to handle large databases. It uses both parallel and distributed models on clusters in 2 main components, Mapper and Reducer, respectively. With regard to pattern mining and the MapReduce framework, the authors in [31] proposed 3 algorithms, using Apriori property to discover the necessary and relevant information. To be used in HUIM, the authors in [29] invented PHUI-growth to be used in the mining of HUIs from big data. As CHUIM research rapidly grows, efficient model development is a necessity for discovering CHUIs in large-scale databases. We refer readers to [29–31] for more in-depth information

on the MapReduce framework and skip an in-depth discussion here in lieu of space considerations in the manuscript.

2.2. Evolutionary Computation. Genetic algorithm (GA) was presented by Holland [32] as the first optimization approach in evolutionary computation. The benefit to use GA is that it is not a trivial task to implement GA for real applications. GA is used to solve the NP-hard question and provides a solution optimally. The idea for GA implementation is to encode the solutions as a chromosome, and each chromosome is represented as an individual in the population. To evaluate the goodness of the chromosome, a fitness function should be predefined in the evolutionary process. Since GA is the fundamental approach in evolutionary computation, many extensions [33, 34] are then developed and studied to enhance its efficiency.

In GA, 3 operations are generally considered to iteratively perform for obtaining a better solution, and they are indicated as (1) mutation, (2) crossover, and (3) selection. For the evolutionary progress of GA, first, each possible solution is then encoded as a chromosome, which can be presented as a string by binary or decimal encoding scheme. The crossover operation is then performed to swap the parts of the chromosomes that can be used to produce the offspring as a new solution for the next generation. The idea of crossover operation is to generate the possible solutions and better convergence in a search space. After that, a mutation operation is then executed to flip some digits of a chromosome, which generates new solutions. The idea of mutation operation is to change parts of a solution randomly, which can increase the diversity of the population and provide a mechanism for escaping from a local optimum. Note that the ratio for running the crossover and mutation is different, and normally, the ratio of crossover operation is higher than that of the ratio of mutation operation. After that, the selection operation is then operated to find the elite solutions for the next round (or generation). This selection mechanism is mostly based on the fitness value. Thus, iterative progress is then performed until the termination condition is achieved. Several criteria can be set to terminate the progress of evolutionary model by (1) the number of iterations is achieved by the predefined the number of generations or (2) the fitness value becomes stable without further big changes; the algorithm is converged. However, in traditional GA-based model, it takes long time to be converged by the 3 generic operations.

Several EC-based approaches were adapted to generic ARM [35], HUIM [13, 36], and high average-utility itemset mining (HAUIM) [37] for knowledge discovery. Qodmanan et al. [35] presented a GA-based model to mine the association rules without minimum support and confidence thresholds. The designed fitness function can produce more interesting and important rules rather than the traditional approaches. Kannimuthu and Premalatha [36] first adapted the GA-based model in HUIM that can discover the set of the HUIs in a limit time. Gunawan et al. [13] presented a BSPO model for mining HUIM without threshold value. Further extensions are then developed in progress to adapt the evolutionary computation (EC) for mining the required information. Song and Huang [37] used the PSO model for revealing the high average-utility itemsets.

2.3. High-Utility Itemset Mining (HUIM). There can be very beneficial reasons to analyse the purchase behaviours of customers in basket-market domains since the revealed information and knowledge will provide the realistic and profitable values of the products to the company, e.g., supermarket or shopping mall. Generic models of association rule mining/frequent itemset mining only take occurrence frequency as the major consideration, which provide the insufficient knowledge to make the efficient decision especially it is not applicable on an item with lower frequency in the database but can bring higher profit than the others, i.e., diamond or caviar. HUIM [16] was presented to take the internal factor (considered as the quantity of the item in the transactions) and external factor (considered as unit profit for the item in any database) to reveal the set of HUIs, which shows an alternative model for making more precise and accurate strategies for decision-making.

Traditional models of HUIM [16] do not hold the DC property; thus, it takes a very huge search space by "combinational explosion" mechanism to reveal the required information. TWU model [14] was presented to build the upper-bound values on the itemsets by holding and maintaining the HTWUIs. This model can hold the TWDC property to solve the limitation of the past HUIM models. Although TWU model is efficient but it still builds the very high upper-bound values on the itemsets; thus, several models were, respectively, presented to mine the set of HUIs and speed up mining performance. The high-utility pattern- (HUP-) tree was developed to keep the required information into a tree structure, which provides good performance than that of the traditional TWU model. Utility-pattern- (UP-) growth and UP-growth+ [17] were then developed to mine the set of HUIs efficiently from the implemented utility-pattern tree. The above algorithms are, however, still based on TWU model to keep the loose upper-bound values on itemsets; thus, the number of discovered candidates in phase 1 is still a lot. To reduce this limitation by having a lot of candidates in phase 2, HUI-Miner [15] was designed and implemented by a linked-list structure named utility-list- (UL-) structure that can avoid the generate-and-test and tree-based models for mining the set of HUIs. It also uses the join operator to generate k-itemsets; thus, the required HUIs at different levels can be found and discovered efficiently. FHM [19] was investigated to build a matrix structure effectively to store the cooccurrence relationships among itemsets that can be used to reduce the search space efficiently since the unpromising candidate itemsets can be early pruned and removed. FIM [38] was then developed and implemented to work on two strategies that can be used to establish the tight upper-bound values on the itemsets; the size of the search space can be reduced greatly. Several works of HUIM are then extensively studied and discussed. Srivastava et al. [39] used the prelarge and fusion models to mine the set of HUIs from wireless sensor networks for the real industry applications. Several approaches and studies are then developed in HUIM, and this research issue has been still developed in progress [9, 40].

Although most of the pattern mining models, e.g., ARM or HUIM, can find the required information for decision-

making, it is sometimes not a trivial task to retrieve the most useful and meaningful information from a huge number of the rules especially for some online decision-making system, i.e., stock market analysis. Thus, it is possible to provide less but meaningful information and knowledge for further decision-making. Closed pattern mining of frequent itemset mining [22, 23] is a good model to find the less but concise patterns as the solution for decision-making. Instead of mining a high number of patterns for decision-making, closed frequent itemset mining can greatly reduce the size of the discovered patterns; thus, it is somehow easier to make the decision in a short time. Closed-pattern mining model was also adapted the concept of HUIM; thus, the CHUI-Miner [27] was presented to find the CHUIs in the databases. Since CHUI-Miner is a one-phase approach; thus, it uses the EU-list model to keep the necessary information for the later mining progress. However, the CHUI-Miner still relies on TWU property to maintain the upper-bound values on the itemsets; it still suffers the limitation of huge search spaces for finding the required patterns; thus, the execution time is costly. Up to now, the state-of-the-art model called CLS-Miner [6] was presented that incorporates the UL-structure EUCS strategy in the mining progress. The EUCS model is very beneficial to reduce the number of 2-itemsets for the further progress; thus, the size of search space can be greatly reduced. Moreover, CLS-Miner applies the efficient strategies to prune the size of the search space as well; thus, the mining performance can be sped up. Up to now, none of the existing models can thus be used to handle the large-scale databases for mining the CHUIs, which is the major task and research issue in this work.

3. Preliminary and Problem Statement

A set of items in the database is denoted as I and defined as $I = \{i_1, i_2, \cdots, i_m\}$. Also assume that a database is denoted as D and defined as $D = \{T_1, T_2, \cdots, T_n\}$. Note that each $T_d \subseteq I$ $(1 \le d \le n)$, and there is n transaction in the database D. Suppose that the quality of an item i_j in a transaction T_j is denoted as $q(i_j, T_d)$, and the unit profit of an item i_j is denoted as $p(i_j)$. Note that both $q(i_j, T_d)$ and $p(i_j)$ are the positive integers. Assume that an itemset is denoted as X such that $X = \{i_1, i_2, \cdots, i_k\}$. The length of X is considered the size of the itemset X, which can be considered as k -itemset $(k = 1, 2, \cdots, m)$. Key definitions of this paper are given as follows.

Definition 1. The utility of an item i_j in a transaction T_d is denoted as $u(i_j, T_d)$ and defined as follows:

$$u(i_j, T_d) = q(i_j, T_d) \times p(i_j), \tag{1}$$

where $q(i_j, T_d)$ is the quantitative value of i_j in T_d and $p(i_j)$ is the unit profit of an item i_j in the unit of the profit table.

Definition 2. The utility of an itemset X in a transaction T_d is denoted as $u(X, T_d)$ and defined as follows:

$$u(X, T_d) = \sum_{i_j \in X} u(i_j, T_d). \tag{2}$$

Definition 3. The utility of an itemset X in a database D is denoted as $u(X)$ and defined as follows:

$$u(X) = \sum_{X \subseteq T_d \wedge T_d \in D} u(X, T_d). \tag{3}$$

Definition 4. The utility of a transaction T_d is denoted as t $u(T_d)$ and defined as follows:

$$tu(T_d) = \sum_{i_j \in T_d} u(i_j, T_d). \tag{4}$$

Definition 5. The total utility of a database D is denoted as $u(D)$ and defined as follows:

$$u(D) = \sum_{T_d \in D} tu(T_d). \tag{5}$$

Definition 6. Suppose an itemset is defined as X, and the minimum utility threshold is set as δ. An itemset is a high-utility itemset (HUI) if it follows the following condition as

$$u(X) \ge \delta \times u(D). \tag{6}$$

Definition 7. Suppose an itemset X is a CHUI. It must have the following conditions as follows: (1) any superset (i.e., Y) of X will not have the same support value such as $\sup(Y) = \sup(X)$ and (2) $u(X)$ is larger than or equal to the minimum utility count. Note that $u(Y)$ is also larger than or equal to the minimum utility count.

For the generic association rule mining or frequent itemset mining, it holds the downward closure property to avoid the "combinational explosion" issue. To increase the mining performance in HUIM, a new property called transaction-weighted downward closure (TWDC) was established by TWU model [14] that can be used and adapted in HUIM to solve the limitation of the generic models.

Definition 8. An itemset is denoted as X, and its transaction-weighted utility is denoted as $\mathrm{twu}(X)$. To calculate the transaction-weighted utility of X, it follows the condition as follows:

$$\mathrm{twu}(X) = \sum_{T_d \in D \wedge X \subseteq T_d} tu(T_d). \tag{7}$$

Current works [14, 17, 19] regarding HUIM applied the TWU model to keep the TWDC property; it also adapts to CHUIM [27] to avoid the problem of "combinational

explosion." In addition, the UL-list-based model [15] and EUCS-based approach [19] are beneficial to efficiently reveal the required high-utility itemsets. For example, UL-list uses the join operator, which is easily to find the $(k+1)$-itemsets level wisely without candidate generation. The EUCS model uses the matrix structure to keep the TWU values of 2-itemsets. Based on the DC and TWDC properties, if a 2-itemset is not a HTWUI, its superset will not be the HTWUI either; the superset of the itemset can be discarded and ignored. Thus, the search space can be reduced efficiently. As we mentioned, the CHUIM can produce a smaller number of useful and meaningful patterns; thus, it is possible to make the decision quickly based on some specific online applications. The generic models [27, 28] of CCCCCCHUIM cannot, however, handle the large-scale and big datasets, which is not applicable in real-life situations and applications. We thus then developed a MapReduce framework that can be used to process the CHUIM in very big and large-scale datasets.

Problem Statement: Suppose a very large transactional database D, and each transaction in D consists of the purchased items with their quantity values. A profit table is assumed as a ptable that keeps the unit profit of the items in the database. Let δ be the minimum utility threshold in the database. The purpose of this paper is aimed at finding the complete set of the CHUI efficiency by the cloud-computing techniques for handling the large-scale datasets.

4. The Developed GA-Based MapReduce Model for CHUIM

We first design a GA-based decomposition model and a 3-tier MapReduce framework for handling large-scale CHUIM in this section. The idea of exploring the decomposition and combining the 3-tier MapReduce is to reduce the search space for finding the required information, which easily is explored by the genetic algorithm (GA). First, the set of transactions D is then partitioned into several groups $G = \{G_1, G_2, \cdots, G_k\}$, in which each group G_i contains several transactions in D, and k is set as the group number in the database. Generally, the groups hold disjoint relationship, in which for every two different groups, it holds the condition as follows:

$$\left(G_i, G_j\right), I(G_i) \cap I\left(G_j\right) = \varnothing, \tag{8}$$

where $I(G_i)$ is the set items of the group G_i and $I(G_j)$ is the set items of the group G_j.

Proposition 9. *Let G be the groups of transactions in the original database D. If the groups in G have no shared items, the set of all relevant frequent itemsets is considered as the unions of the full groups' frequent itemsets. We thus can note that*

$$F = \left\{ \bigcup_{i=1}^{k} F_i \right\}, \tag{9}$$

where F_i is considered as a set of the relevant frequent itemsets of the group G_i.

Proof. Consider $\forall (i, j) \in [1 \cdots k]^2, I(G_i) \cap I(G_j) = \varnothing$, we can obtain that $\forall i \in [1 \cdots k]$: $F_i = \{p \mid \sup(D, I, p) \geq \min\_\sup\}$. The support of the pattern p is examined by checking all transactions in D. Considering a pattern p exists in $I(G_i)$, i.e., $p \subseteq I(G_i) \Rightarrow \forall e \in p, e \in I(G_i) \Rightarrow \forall e \in p, e \notin I(G_j), (\forall j \in [1 \cdots k], \vee j \neq i) \Rightarrow p \notin I(G_j) \Rightarrow F_i = \{p \mid \sup(G_i, I(G_i), p) \geq \min\_\sup\} \Rightarrow F = \{\bigcup_{i=1}^{k} F_i\}$.□□ □

The proposition above clearly shows that transactions that are in D must follow certain conditions above, from which the dependent groups can be fully revealed. Thus, relevant frequent itemsets can be identified using pattern mining approaches in groups. However, this is not a realistic scenario, and the objective is to decrease the number of items shared by the separated groups. Existing work [5] identified that k-means [41] and DBSCAN [42] can obtain a good performance of transaction decomposition, and k-means showed better results than that of the DBSCAN. Thus, a k-means model is used in the designed framework for transaction decomposition that can group highly relevant transactions in the same group. After that, a GA-based MapReduce- (GMR-) Miner algorithm that consists of GA and 3-tier MapReduce framework for mining the closed patterns with high utilization is then presented. Three phases in the designed framework regarding different MapReduce tasks are described below.

4.1. Exploration. After dividing the transactions into several groups, each Mapper is fed with a partition. The framework for MapReduce is applied in this step for the exploration of any and all promising items which may be CHUI in addition to their supersets. Any unpromising itemsets can easily be discarded in this step to make good mining progress due to the design properties which can be stated as follows.

Property 10. We can say that if there exists a known pattern t that clearly is or can be defined as a frequent pattern, it can be defined as a frequent itemset in one part.

Proof. Let a database D being split into n parts such that $\{D_1, D_2, \cdots, D_n\}$; the total frequency of each part is calculated as $\{|D_1|, |D_2|, \cdots, |D_n|\}$. Assume that the minimum support threshold is considered as δ in the database, and t is considered as a frequent pattern in D. We then can obtain the following situation as follows:

$$s(t) \geq \delta \times |D_i|. \tag{10}$$

The counter-evidence, $\{s_1, s_2, \cdots, s_n\}$, is used to show the support value of an itemset (pattern) t of each part. Obviously, t is not considered as the frequent itemset in the entire part such that $s_1 < \delta \times |D_1|, s_2 < \delta \times |D_2|, \cdots, s_n < \delta \times |D_n|$. Then, $s(t) = \sum_{i=1}^{n} |D_i|$ is different to the above

definition. This, we can prove that the correctness is held by this property.□□ □

Based on the developed Property 10, it is then studied and extended to the designed MapReduce model. Thus, the integrity of the mined information is then ensured. According to the DC property used in the Apriori algorithm, Property 11 is studied and extended from Property 10 to ensure that the supersets can satisfy the condition. The definition is then given as follows.

Property 11. Suppose two itemsets t and t' hold the situation such as $t \subseteq t'$. Thus, in the database D, we can ensure that $s(t, D) \geq s(t', D)$ maintains the correctness.

Based on Property 11, if a support of an itemset t is less than the minimum support threshold (count, $\delta \times |D|$), it is not treated as a frequent itemset, neither its supersets. That is, it does not affect the final results if t is then early removed. In the proposed paper, each Mapper acquires a database partition. Thus, the pair of <key, value> for an itemset with its support value (or called frequency) is output in a certain partition to the Reducer. Following that, a GA-based technique is used to investigate the possible search space for the next Reducer phase. All frequent itemsets are treated as individuals in the first population, and then, the unsatisfied itemsets are removed to efficiently minimize the search space for later processes. This GA-based technique can significantly cut computational costs by avoiding the need to explore the whole search space. Following that, all promising frequent itemsets are inspected to determine the complete closed frequent itemsets [22, 23] by the next MapReduce framework to reveal the satisfied CHUIs.

To be concluded, the initial MapReduce divides the clustered dataset into numerous parts (or called partitions), which are subsequently processed independently by each Mapper. The GA model then generates a search space for prospective candidates that can be used to reduce the size of the search space. Following that, all satisfied frequent itemsets are mined and revealed, and unsatisfied frequent itemsets are deleted here. Once again, only satisfied CHUIs will be sent to the subsequent MapReduce model for revealing the set of CHUIs. The following is the description of the exploitation phase.

4.2. Exploitation. The exploitation phase begins with the usage of current CHUIM models (e.g., CLS-Miner [28]) to mine the CHUIs for each partition. Given that mining the set of CHUIs in the whole dataset is not straightforward, the second MapReduce is executed in parallel with the partial, tiny, and numerous sets from promising itemsets from the initial MapReduce on each node. Due to the fact that each node requires less memory, the MapReduce architecture is capable of running a large database on a single machine. The candidate's utility is then explored for each node in order to determine the progress of the exploitation mining. The horizontal structure known as tidset is used to store the transaction ID and its associated frequent itemsets. Due to the efficient tidset structure, it is simple to calculate the frequencies of the itemsets in the mining progress; thus, the computational cost can be greatly minimized and the performance can be greatly improved.

Additionally, a straightforward load balancing strategy is used to divide the transactions into the second MapReduce tasks based on their sizes before performing the second MapReduce. The computational cost of the exploitation process can thus be decreased. The number of produced tasks should correspond to the number of Mappers. The workload of each node is determined by the amount of promising itemsets in a transaction, and then, the transaction is assigned to the node with the least workload, which is capable of evenly distributing the computation among the nodes. When compared to the serialization model, this technique can significantly lower processing costs. The load balancing equation is given as follows.

$$\mathrm{WL}_i = \mathrm{WL}_i + \mathrm{Num}, \tag{11}$$

where WL_i is the workload of node i, and Num is the number of patterns derived from the first MapReduce of the performed transaction. The CLS-Miner [28] is applied here to mine the set of local CHUIs at each partition D_i. The local CHUI is then output from each Mapper, and the result is a pair of $\{\text{pattern}, (\text{utility}\,; p_i)\}$.

The Mapper stage first executes the CLS-Miner to mine the set of CHUIs within the partition and then assigns the local CHUIs with the same key (or itemset/pattern) to the same Reducer. It is possible to calculate the partial total utility in a partition; the local CHUI can be recognized if its utility value is not smaller than the sum of the partial total utility in the partition. As a result, the CHUI that has been satisfied is output to the result file; otherwise, the Reducers output the key-value pair that will be used later in the generation of the candidate set. Following that, all candidates (possible patterns) and the tidset are required for the next-generation phase, which is completed during the second MapReduce phase. Crossover and mutation procedures are done on the second MapReduce framework to produce the possible candidates for the actual CHUIs between the Mapper and Reducer of each partition.

In this phase, each MapReduce component considers only one cluster of transactions. This allows to highly reduce exploring the solution space. At the same time, the candidate patterns have been calculated for their utilities of each node. Therefore, by using a developed tidset structure, the calculated utility can be used to speed up the checking process.

4.3. Integration. The purpose of this stage is to catch any patterns that have been missed in the local clusters due to mining progress. It takes into account both shared and clusters during the exploration and exploitation processes. This enables the discovery of all associated CHUIs across the whole database. From the shared items, potential candidate CHUIs are established initially. It is then investigated to find the significance of each generated pattern over the entire database by utilizing the integration function. The designed framework proposes an aggregation function, which is the

Input: D, a quantitative database; *ptable*, a profit table of all items; δ , a minimum utility count.
Output: a set of discovered closed high-utility itemsets (CHUIs)
perform k-means to cluster D as $(p_1, p_2, \cdots, p_n)$.
perform **exploration** function {
for each p_k {
  set key-value pair as (*tid*, t-itemset).
  for each t-itemset {
    calculate sup (t).
    }
  write$(t,<$ sup $(t), p_k >)$.
  }
  for each t in p_k {
    sup (t) = sup (t) + <sup $(t), p_k >$.
    }
  write$(t,$ sup $(t))$.
}
perform **exploitation** {
build *tidset*.
for each p_k {
  set key-value pair as (*tid*, t-itemset).
  for each t-itemset {
    calculate $u(t)$ by CLS-Miner.
    }
  write$(t,<u(t), p_k >)$
  }
  for each t in p_k {
    $u(t) = u(t) + <u(t), p_k >$.
    }
  write$(t, u(t))$.
}
perform **integration** {
project t-itemset as utility-list (UL).
build EUCS of 2-itemsets.
for each C in t {
  check Cq-itemset.
  if C appears in *tidset* and *tid* == key {
    write a pair $(C, lu(C))$.
    else {
    write a pair $(C, lu(C))$.
    }
  }
}
for each C in t {
$gu(C) = gu(C) + lu(C)$.
  if $gu(C) >= \delta \times u(D)$ {
    write$(C,$ gu$(C))$.
  }
}
}

Algorithm 1: The designed GMR-Miner algorithm.

sum of local support for shared patterns across all clusters, to be used as an integration function. Afterwards, the relevant CHUIs of the shared items are concatenated with the relevant CHUIs of the local clusters to derive the globally relevant patterns across the entire transaction database. Additionally, the tidset generated by the second MapReduce is used to decrease the computation required to mine the patterns of each node. Additionally, the utility-list structure

Table 1: The parameters of the used databases.

Dataset	$\|D\|$	$\|I\|$	C	MaxLen
SIGN	730	267	52	94
Leviathan	5,834	9,025	33.8	100
MSNBC	31,790	17	13.3	100
BMS	59,601	197	2.5	267

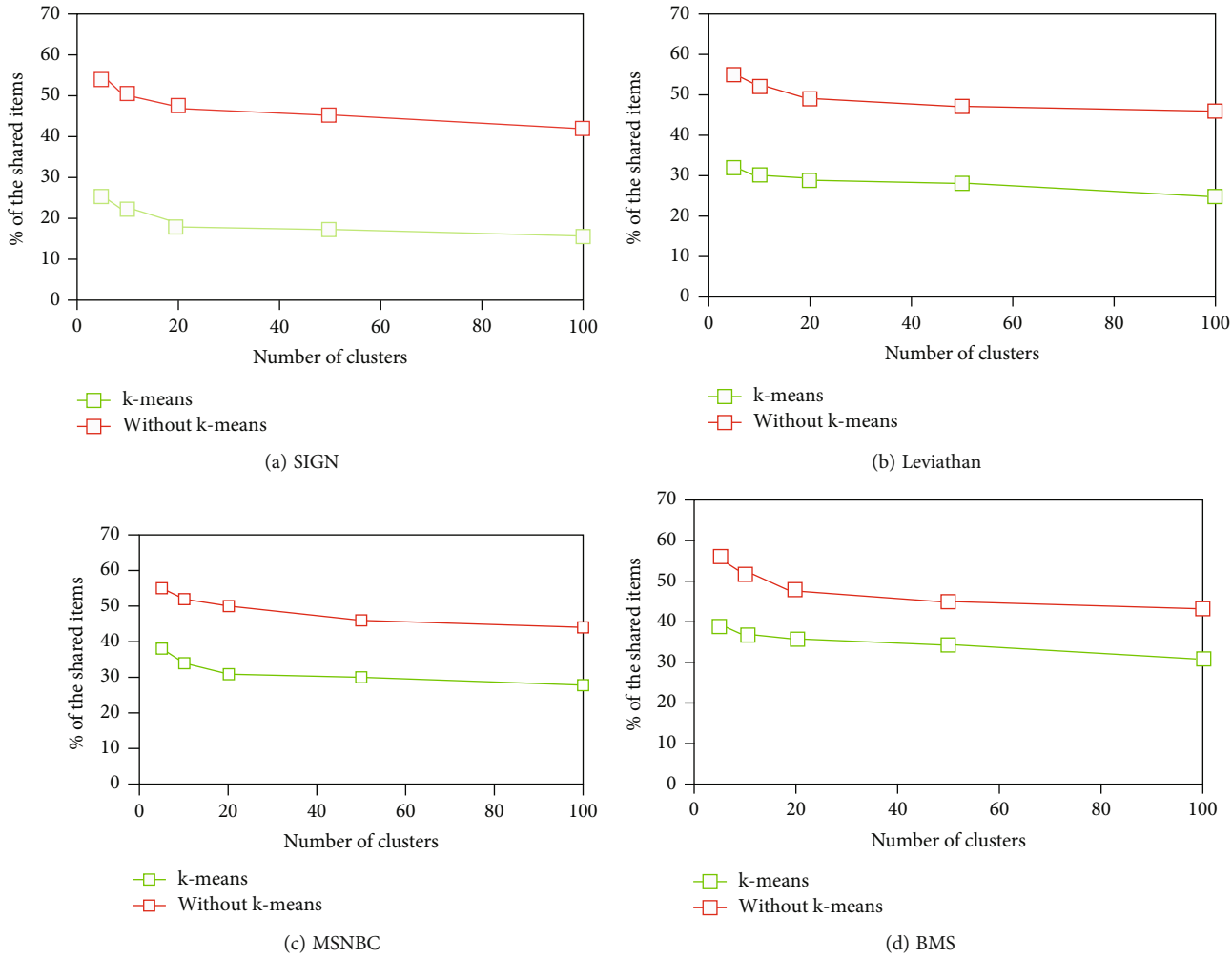


Figure 1: The decomposition clustering quality.

and EUCS are constructed to hold the data required for the calculation, and the computational cost is lowered as a result of these two structures. Additional information on utility-list and EUCS is available in [28].

The third MapReduce framework can then be used to determine and identify global patterns about CHUIs using the set of candidate patterns (local CHUIs) and the tidset. The genetic algorithm's selection operator is used in this phase to retain only the relevant patterns for the next generation, and the fitness (utility) of each candidate pattern is calculated. Each Mapper stage converts the information in the itemset into a utility-list and then determines the local utility of all itemsets in the candidate set. Besides, the EUCS is then applied here to reduce computation if the investigated itemset does not meet the needs. If an investigated itemset can be found from tidset by using its transaction ID, it shows that the utility of the itemset was determined before in the second MapReduce stage; the Mapper here then delivers a pair value of regarding pattern and its utility such as (pattern; utility) for the next Reducer phase. Otherwise, the utility of the pattern can be thus determined by the utility-list structure and a pair value is then output as the result. According to three strategies here such as EUCS, tidset, and utility-list, the mining progress can be sped up, and the computational cost is then reduced for finding the global itemsets with their utility values in the entire database. The Reducer stage here is considered to sum up the utilities of the investigated pattern, and if this value is larger than the $\delta \times u(D)$ in the Reducer stage, it is the globally CHUI and will be released as the final output of the designed framework. Detailed progress of the designed framework is then shown in Algorithm 1.

5. Experimental Evaluation

In the experiments, four realistic databases [43] are then used in this paper to state the performance of the developed GMR-Miner approach compared to the state-of-the-art CLS-Miner [28] model in terms of runtime, memory usage, and scalability under a varied number of nodes in the developed 3-tier MapReduce framework. Note that the developed MapReduce is then deployed in Spark since Spark provides a higher capability to handle the large-scale databases. The properties of 4 conducted databases are then described in Table 1. Here, $|D|$ is the number of database size, which showed the number of transactions in the database. $|I|$

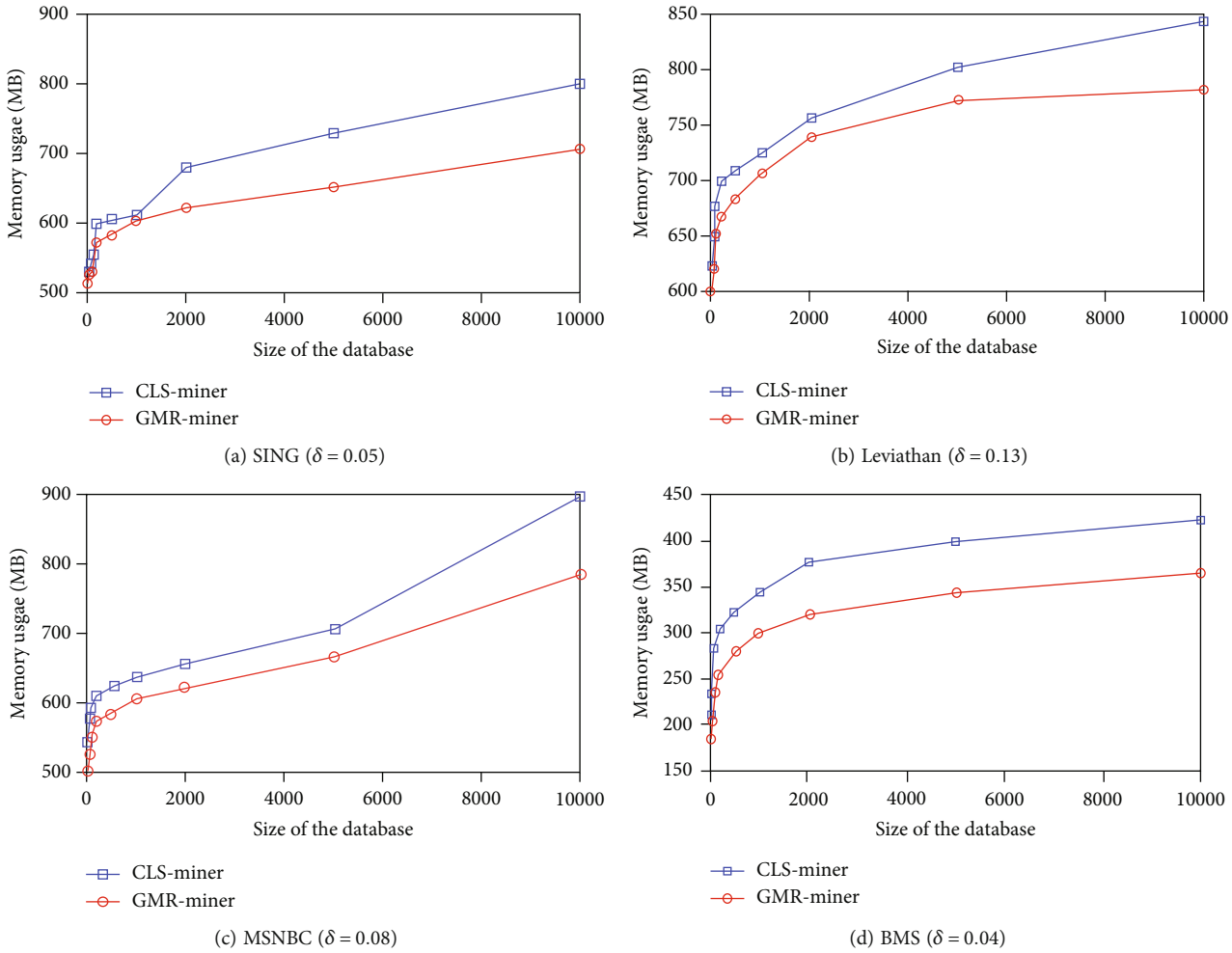
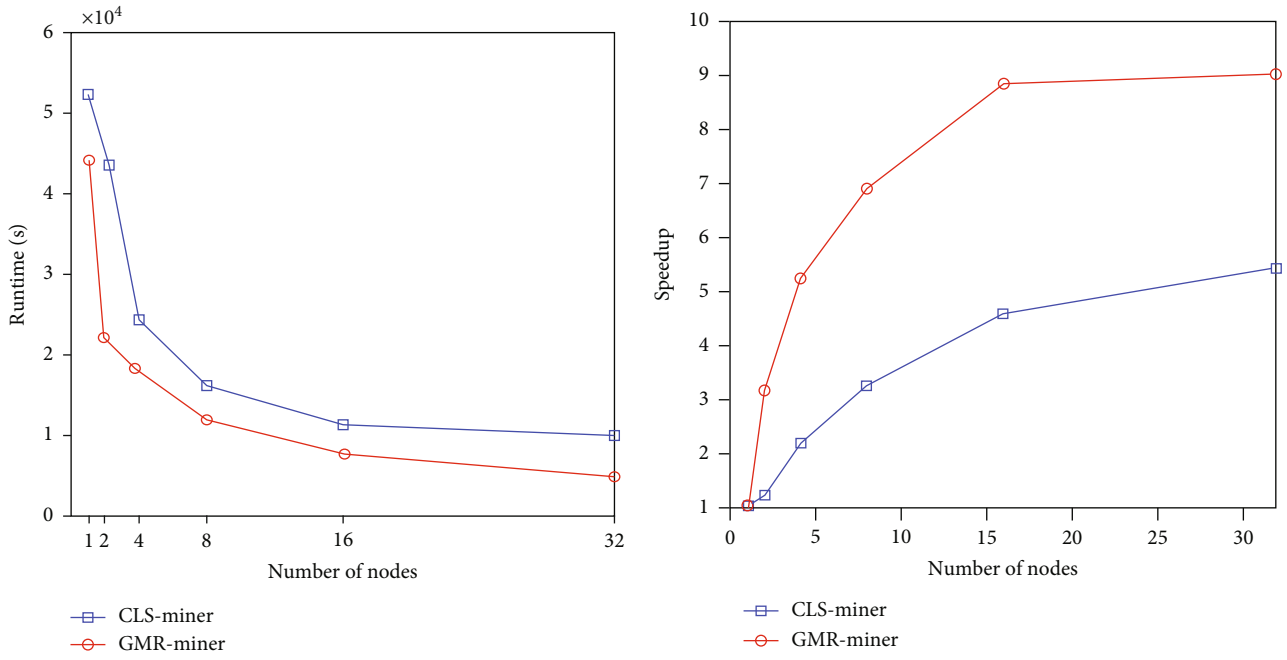


Figure 2: Memory usage of the compared algorithms.

indicated the number of distinct items in the database. C showed the average number of items in a transaction, and MaxLen is the maximum size of a transaction in the database. The used databases in Table 1 are then enlarged and duplicated by various numbers (e.g., 1, 20, 50, 100, 200, 500, 1,000, 2,000, 5,000, and 10,000) for the later performance evaluation.

5.1. Quality of Clustering. Figure 1 shows the quality evaluation of the returned clusters by using the k-means and the intuitive clustering algorithm on the four datasets used in the experiments. The intuitive clustering divides the transactions into k-clusters randomly without any processing. In the conducted experiments, the quality of the returned clusters is then decided by the % of the shared items in clusters, and the object here is to lower the value. We also set the number of clusters in the experiments from 1 to 100; thus, the % for the shared items is then reduced for the evaluation with and without k-means approach. However, there is a large difference between k-means and intuitive algorithms in all cases. For instance, by using k-means to split the transactions, the percentage of shared items does not exceed 40%. However, without using the k-means, the percentage of shared items reaches 60%. With the further explanations by the property of k-means model, it finds the centroid point based on the similarity equation; the intuitive idea only processes the points by randomness operations. Overall, these experiments clearly showed the benefit of k-means in data decomposition. Thus, we can observe that the k-means model adapted in this MapReduce framework is useful and effective to mine the CHUIs in large-scale databases.

5.2. Memory Usage. To demonstrate the usability of the developed MapReduce model, the results are carried and compared to the CLS-Miner [28] in terms of memory usage, which are shown in Figure 2. By varying the size of the database, it can be seen that the developed GMR-Miner outperforms CLS-Miner in all cases. For instance, only 350 MB is needed by the GMR-Miner to deal with 10,000 times of BMS data. However, 420 MB is needed by the CLS-Miner to handle the same data. These results are reached due to the decomposition step, where each cluster contains similar transactions, and also the intelligent operators of the genetic algorithm where it accurately explores the possible solution space. Thus, less memory usage is then required by the developed GMR-Miner compared to CLS-Miner algorithm.



(a) BMS × 1000 (b) BMS × 1000

Figure 3: Scalability results in terms of runtime and speedup.

Table 2: Clustering quality versus pattern mining accuracy.

Data	% of the shared items	% of the relevant patters
SIGN	40	79
	35	88
	30	96
Leviathan	40	77
	35	83
	30	95
MSNBC	40	72
	35	88
	30	97
BMS	40	84
	35	86
	30	91

5.3. Scalability. To show that the designed GMR-Miner achieves good robustness and applicable in real applications for the large-scale scenario, the scalability on a big dataset is illustrated in Figure 3. Here, we duplicated the BMS dataset 1,000 times for scalability evaluation under a varied number of nodes from 1 to 32. The results showed that the developed GMR-Miner outperforms the CLS-Miner in terms of runtime and speedup under a varied number of nodes, where a high gap between the two approaches is observed. For instance, with 32 nodes, the speedup of the GMR-Miner is 9 for handling 1,000 times of BMS data. However, the speed up of the CLS-Miner is only 5 to handle the same data and with the same number of nodes. This result confirmed the usefulness of genetic algorithms and decomposition for discovering CHUIs in big and large-scale datasets. In general, the developed model can easily process the very big and large databases for mining the required CHUIs, which is very suitable and appropriate for the market engineering.

5.4. Clustering Quality vs. Pattern Mining Accuracy. Table 2 presents the quality of the pattern mining process with varying on the clustering quality using the four data (SIGN, Leviathan, MSNBC, and BMS). By varying the quality of the clustering detected by the % of the shared items in the clusters from 40% to 30%, the accuracy of the pattern mining solution increases from 70% to 90% for all the databases used in the experiments. This result is reached thanks to the low dependency among clusters, where the mining process may be applied differently on each cluster of transactions.

6. Conclusion and Discussion

Mining high-profitable and concise patterns in IoT environments is not a trivial task since the collected data is usually a large-scale dataset. Past studies of mining CHUIs cannot handle (1) large-scale dataset and (2) mining the required information in a limited time. In this paper, we used a 3-tier MapReduce framework deployed in Spark for efficiently mining the closed patterns with high utilization (or a.k.a. CHUIs). To better explore the possible and potential candidates instead of the entire search space, the genetic algorithm (GA) is also utilized in the designed model for better pattern exploration progress. Experiments are then showed that the designed GMR-Miner outperforms the CLS-Miner in terms of execution time, memory, and scalability regarding a different number of nodes. In the future, a better data

structure can be deployed instead of a utility-list structure for obtaining better performance, and the incremental model can also be investigated and explored as a further research topic to handle the issue of dynamic data mining. In addition, to find the sufficient and satisfied solutions in a limit time, other algorithms such as PSO or ACO in evolutionary computation can also be explored and studied as the further extension.

Data Availability

The data used to support the findings of this study have been deposited in the SPMF repository (doi:10.1007/978-3-319-46131-1_8).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Western Norway University of Applied Sciences, Norway, provides partial funding support for the work carried out in this paper.

References

[1] M. J. Zaki and C. J. Hsiao, "Efficient algorithms for mining closed itemsets and their lattice structure," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 462–478, 2005.

[2] B. Lin, F. Zhu, J. Zhang et al., "A time-driven data placement strategy for a scientific workflow combining edge computing and cloud computing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4254–4265, 2019.

[3] Y. Qu and N. Xiong, "RFH: a resilient, fault-tolerant and high-efficient replication algorithm for distributed cloud storage," in *2012 41st International Conference on Parallel Processing*, pp. 520–529, Pittsburgh, PA, USA, 2012.

[4] R. Agrawal, T. Imielinski, and A. N. Swami, "Database mining: a performance perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, no. 6, pp. 914–925, 1993.

[5] A. Belhadi, Y. Djenouri, J. C. W. Lin, and A. Cano, "A general-purpose distributed pattern mining system," *Applied Intelligence*, vol. 50, no. 9, pp. 2647–2662, 2020.

[6] G. Grahne and J. Zhu, "Fast algorithms for frequent itemset mining using FP-trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 10, pp. 1347–1362, 2005.

[7] R. U. Kiran, A. Anirudh, C. Saideep, M. Toyoda, P. K. Reddy, and M. Kitsuregawa, "Finding periodic-frequent patterns in temporal databases using periodic summaries," *Data Science and Pattern Recognition*, vol. 3, no. 2, pp. 24–46, 2019.

[8] H. Si, J. Zhou, Z. Chen et al., "Association rules mining among interests and applications for users on social networks," *IEEE Access*, vol. 7, pp. 116014–116026, 2019.

[9] U. Yun, H. Ryang, and K. H. Ryu, "High utility itemset mining with techniques for reducing overestimated utilities and pruning candidates," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3861–3878, 2014.

[10] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, 2004.

[11] R. Chan, Q. Yang, and Y. D. Shen, "Mining high utility itemsets," in *IEEE International Conference on Data Mining*, pp. 19–26, Melbourne, FL, USA, 2003.

[12] W. Gan, J. C. W. Lin, P. Fournier-Viger, H. C. Chao, V. Tseng, and P. S. Yu, "A survey of utility-oriented pattern mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 1306–1327, 2021.

[13] R. Gunawan, E. Winarko, and R. Pulungan, "A BPSO-based method for high-utility itemset mining without minimum utility threshold," *Knowledge-Based Systems*, vol. 190, article 105164, 2020.

[14] Y. Liu, W. Liao, and A. N. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2005*, T. B. Ho, D. Cheung, and H. Liu, Eds., vol. 3518 of Lecture Notes in Computer Science, pp. 689–695, Springer, Berlin, Heidelberg, 2005.

[15] M. Liu and J. Qu, "Mining high utility itemsets without candidate generation," in *ACM International Conference on Information and Knowledge Management*, pp. 55–64, Maui, HI, USA, 2012.

[16] H. Yao, H. J. Hamilton, and C. J. Butz, "A foundational approach to mining itemset utilities from databases," in *SIAM International Conference on Data Mining*, pp. 482–486, Lake Buena Vista, Florida, US, 2004.

[17] V. S. Tseng, B. Shie, C. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," *IEEE Transactions Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1772–1786, 2013.

[18] J. C. W. Lin, T. Hong, and W. Lu, "An effective tree structure for mining high utility itemsets," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7419–7424, 2011.

[19] P. Fournier-Viger, C. W. Wu, S. Zida, and V. S. Tseng, "FHM: faster high-utility itemset mining using estimated utility co-occurrence pruning," in *Foundations of Intelligent Systems. ISMIS 2014*, T. Andreasen, H. Christiansen, J. C. Cubero, and Z. W. Raś, Eds., vol. 8502 of Lecture Notes in Computer Science, pp. 83–92, Springer, Cham, 2014.

[20] J. Liu, K. Wang, and B. C. M. Fung, "Direct discovery of high utility itemsets without candidate generation," in *2012 IEEE 12th International Conference on Data Mining*, pp. 984–989, Brussels, Belgium, 2012.

[21] C. Yin, S. Zhang, J. Wang, and N. N. Xiong, "Anomaly detection based on convolutional recurrent autoencoder for IoT time series," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 21, no. 14, pp. 15626–15634, 2020.

[22] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient mining of association rules using closed itemset lattices," *Information Systems*, vol. 24, no. 1, pp. 25–46, 1999.

[23] C. Lucchese, S. Orlando, and R. Perego, "Fast and memory efficient mining of frequent closed itemsets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 21–36, 2006.

[24] B. Vo, L. T. T. Nguyen, N. Bui, T. D. D. Nguyen, V. N. Huynh, and T. P. Hong, "An efficient method for mining closed potential high-utility itemsets," *IEEE Access*, vol. 8, pp. 31813–31822, 2020.

[25] T. Wei, B. Wang, Y. Zhang, K. Hu, Y. Yao, and H. Liu, "FCHUIM: efficient frequent and closed high-utility itemsets mining," *IEEE Access*, vol. 8, pp. 109928–109939, 2020.

[26] V. S. Tseng, C. W. Wu, P. Fournier-Viger, and P. S. Yu, "Efficient algorithms for mining the concise and lossless representation of high utility itemsets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 726–739, 2015.

[27] C. W. Wu, P. Fournier-Viger, J. Y. Gu, and V. S. Tseng, "Mining closed+ high utility itemsets without candidate generation," in *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 187–194, Tainan, Taiwan, 2015.

[28] T. L. Dam, K. Li, P. Fournier-Viger, and Q. H. Duong, "CLS-Miner: efficient and effective closed high-utility itemset mining," *Frontiers of Computer Science*, vol. 13, no. 2, pp. 357–381, 2019.

[29] Y. C. Lin, C. W. Wu, and V. S. Tseng, "Mining high utility itemsets in big data," in *Advances in Knowledge Discovery and Data Mining. PAKDD 2015*, T. Cao, E. P. Lim, Z. H. Zhou, T. B. Ho, D. Cheung, and H. Motoda, Eds., vol. 9078 of Lecture Notes in Computer Science, pp. 649–661, Springer, Cham, 2015.

[30] J. Dean and S. Ghemawat, "MapReduce," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[31] M. Y. Lin, P. Y. Lee, and S. C. Hsueh, "Apriori-based frequent itemset mining algorithms on MapReduce," in *The International Conference on Ubiquitous Information Management and Communication*, pp. 1–8, Kuala Lumpur, Malaysia, 2012.

[32] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, 1992.

[33] K. Elbaz, S. L. Shen, A. Zhou, D. J. Yuan, and Y. S. Xu, "Optimization of EPB shield performance with adaptive neuro-fuzzy inference system and genetic algorithm," *Applied Sciences*, vol. 9, no. 4, pp. 780–797, 2019.

[34] R. Guha, M. Ghosh, S. Kapri et al., "Deluge based genetic algorithm for feature selection," *Evolutionary Intelligence*, vol. 14, pp. 357–367, 2021.

[35] H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Systems with Applications*, vol. 38, no. 1, pp. 288–298, 2011.

[36] S. Kannimuthu and K. Premalatha, "Discovery of high utility itemsets using genetic algorithm with ranked mutation," *Applied Artificial Intelligence*, vol. 28, no. 4, pp. 337–359, 2014.

[37] W. Song and C. Huang, "Mining high average-utility itemsets based on particle swarm optimization," *Data Science and Pattern Recognition*, vol. 4, no. 2, pp. 19–32, 2020.

[38] S. Zida, P. Fournier-Viger, J. C. W. Lin, C. W. Wu, and V. S. Tseng, "EFIM: a fast and memory efficient algorithm for high-utility itemset mining," *Knowledge and Information Systems*, vol. 51, no. 2, pp. 595–625, 2017.

[39] G. Srivastava, J. C. W. Lin, M. Pirouz, Y. Li, and U. Yun, "A pre-large weighted-fusion system of sensed high-utility patterns," *IEEE Sensors Journal*, 2021.

[40] C. Zhang, G. Almpanidis, W. Wang, and C. Liu, "An empirical evaluation of high utility itemset mining algorithms," *Expert Systems with Applications*, vol. 101, pp. 91–115, 2018.

[41] P. Franti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognition*, vol. 93, pp. 95–112, 2019.

[42] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited," *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1–21, 2017.

[43] P. Fournier-Viger, J. C. W. Lin, A. Gomariz et al., "The SPMF open-source data mining library version 2," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016*, B. Berendt, Ed., vol. 9853 of Lecture Notes in Computer Science, pp. 36–40, Springer, Cham, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9948543, 18 pages
https://doi.org/10.1155/2021/9948543

Research Article

Location Privacy Protection Scheme for LBS in IoT

Hongtao Li,[1,2] **Xingsi Xue**,[3] **Zhiying Li**,[1] **Long Li**,[4] **and Jinbo Xiong**[5]

[1]*College of Mathematics and Computer Science, Shanxi Normal University, Linfen 041000, China*
[2]*Fujian Provincial Key Laboratory of Network Security and Cryptology, Fujian Normal University, Fuzhou 350007, China*
[3]*School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China*
[4]*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China*
[5]*College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350118, China*

Correspondence should be addressed to Long Li; lilong@guet.edu.cn

Received 30 March 2021; Revised 19 July 2021; Accepted 2 August 2021; Published 17 August 2021

Academic Editor: Cong Pu

The widespread use of Internet of Things (IoT) technology has promoted location-based service (LBS) applications. Users can enjoy various conveniences brought by LBS by providing location information to LBS. However, it also brings potential privacy threats to location information. Location data that contains private information is often transmitted among IoT networks in LBS, and such privacy information should be protected. In order to solve the problem of location privacy leakage in LBS, a location privacy protection scheme based on k-anonymity is proposed in this paper, in which the Geohash coding model and Voronoi graph are used as grid division principles. We adopt the client-server-to-user (CS2U) model to protect the user's location data on the client side and the server side, respectively. On the client side, the Geohash algorithm is proposed, which converts the user's location coordinates into a Geohash code of the corresponding length. On the server side, the Geohash code generated by the user is inserted into the prefix tree, the prefix tree is used to find the nearest neighbors according to the characteristics of the coded similar prefixes, and the Voronoi diagram is used to divide the area units to complete the pruning. Then, using the Geohash coding model and the Voronoi diagram grid division principle, the G-V anonymity algorithm is proposed to find k neighbors in an anonymous area so that the user's location data meets the k-anonymity requirement in the area unit, thereby achieving anonymity protection of location privacy. Theoretical analysis and experimental results show that our method is effective in terms of privacy and data quality while reducing the time of data anonymity.

1. Introduction

With the rapid development of the Internet of Things (IoT), mobile computing, GPS, and wireless communication technology, location-based service (LBS) has been widely used in many important fields [1–4]. As the core of the IoT, sensors enable the Internet of Things to realize intelligent perception, object recognition, information collection, and other functions [5, 6]. IoT devices form the backbone of the LBS or LBS applications. Users can enjoy the convenience of location service applications, such as shopping, travel, and accommodation. However, when enjoying the convenience of LBS, users must provide their own location information to LBS servers or IoT devices, which may lead to the disclosure of users' location privacy [7, 8]. Therefore, privacy protection of users' location has become the focus of research in LBS.

At present, domestic and foreign researchers have conducted a large number of studies on location privacy protection and proposed a variety of solutions to the privacy protection problems in LBS, such as location privacy protection technology based on interference, location privacy protection technology based on encryption, and k-anonymity.

The privacy protection technology based on interference mainly uses false information and redundant information to interfere with the attacker's stealing of user information. According to the different user information (identity information and location information), privacy protection technology based on interference can be divided into pseudonym technology and false location technology. The

pseudonym technology hides the real identity of the user by assigning an untraceable identifier to the user, and the user uses the identifier to replace his own identity information for inquiries. False location technology uses false location or adds redundant location information to interfere with the user's location information when the user submits query information.

The location privacy protection technology based on encryption encrypts the user's location and points of interest and then searches or calculates in the ciphertext space, while the attacker cannot obtain the user's location and the specific content of the query. Two typical location privacy protection technologies based on encryption are location privacy protection technology based on private information retrieval (PIR) and location privacy protection technology based on homomorphic encryption.

k-anonymity is a technology proposed by Samarati and Sweeney in 1998. This technology can ensure that each individual record stored in the release dataset cannot be distinguished from other $k-1$ individuals for sensitive attributes so that the probability of a specific individual being found is $1/k$; namely, the k-anonymity mechanism requires at least k records of the same quasi-identifier, so observers cannot connect records through the quasi-identifier. k-anonymity is divided into centralized k-anonymity and k-anonymity under the P2P structure.

Applications based on location-based services bring a lot of convenience to people's lives, but at the same time, it also brings severe challenges to users' privacy and security. When users query information from LBS servers, they need to send personal identity, location, interests, and other information to LBS servers. If this information is leaked by untrusted or malicious LBS servers, the attackers can not only link the user's identity with location and interests but also infer more user private information. Therefore, location privacy protection in LBS is becoming more and more important and has been attached great importance to relevant fields.

The filling curve of Geohash encoding is Peano. The Peano curve was discovered by the Italian mathematician Piano. This curve can fill the space, but it has the shortcoming of sudden change. The commonly used method to solve this problem is to calculate the surrounding 8 areas or fill them with the Hilbert curve space. Because it is in the environment of the road network, this article uses the Voronoi diagram to divide the road network. The use of Voronoi can solve the defect problem of Geohash Base32. Based on Geohash coding and the Voronoi diagram, this paper proposes a road network-oriented location privacy protection method (G-V anonymity algorithm). It can ensure privacy protection, improve the quality of service and the availability of published data, and reduce the time of data anonymity. The main contributions of this paper are as follows:

(1) In the client, this paper proposes the Geohash algorithm, which converts the user's position coordinates into a Geohash code of the corresponding length. The Geohash algorithm converts two-dimensional longitude and latitude into strings, and each string represents a rectangular region. In other words, all the points (longitude and latitude coordinates) in this rectangular area share the same Geohash string, which can protect privacy and make caching easier

(2) On the server side, the user-generated Geohash code is inserted into the prefix tree, the prefix tree is used to find the nearest neighbors according to the characteristics of the coded similar prefixes, and the Voronoi diagram is used to divide the area units to complete the pruning. The advantages of the prefix tree are as follows: use the common prefix of the string to reduce the query time, minimize the unnecessary string comparison, and the query efficiency is higher than the hash tree

(3) Based on the Geohash coding model and Voronoi diagram grid generation principle, this paper proposes the G-V anonymity algorithm, which can find k neighbors in the anonymous area and make the user's location data meet the k-anonymity requirement in the area unit, so as to protect the location privacy. When the number of users is scarce, a corresponding number of dummy elements are produced to meet the k-anonymity requirement and realize location privacy protection

(4) A comprehensive theoretical and experimental analysis is carried out on the proposed method. The experimental results show that the algorithm has high service quality and data availability while completing privacy protection and at the same time has a short data anonymity time

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 introduces Definitions, Geohash Code, Voronoi Diagram, LBS Framework Based on IoT, System Architecture, A Practical Application Scenario, and Attack Model. Section 4 introduces the Geohash algorithm and G-V anonymity algorithm proposed in this paper. Section 5 conducts experiments on the G-V anonymity algorithm in terms of the anonymous success rate, degree of privacy protection, algorithm running time, and antiattack ability of the algorithm. Section 6 is the conclusion of this paper.

2. Related Works

As LBS privacy has become the focus of research, more and more scholars have paid close attention to LBS privacy protection methods. At present, the main methods of location privacy protection include location privacy protection technology based on interference, location privacy protection technology based on encryption, and k-anonymity.

The privacy protection technology based on interference mainly uses false information and redundant information to interfere with the attacker's stealing of user information [9–12]. In Reference [13], they proposed a dynamic pseudonymous-based multiple mix-zones authentication protocol that only requires mobile vehicles to communicate with the reported server for registration and dynamic

pseudonym change. Furthermore, a mechanism is proposed to provide users with dynamic pseudonyms, named as base pseudonyms and short-time pseudonyms, to achieve users' privacy. In Reference [14], a new privacy-preserving solution for pseudonym on-road on-demand refilling is proposed where the vehicle anonymously authenticates itself to the regional authority subsidiary of the central trusted authority to request a new pseudonym pool. The logical demonstration proved that this privacy-preserving authentication is assured. In Reference [15], a novel dynamic mixed zone establishment scheme is proposed to protect the location privacy of autonomous vehicles in the convoy driving context. Compared with the scheme to protect location privacy in the traditional vehicular network, the proposed scheme has less overhead and a higher level of security.

Location privacy protection technology based on encryption can be divided into two categories: location privacy protection technology based on private information retrieval (PIR) [16, 17] and location privacy protection technology based on homomorphic encryption [18, 19]. In Reference [20], they proposed a new method to adjust the vehicle speed which reduces the vehicle delay that suffers from the network gap problem. It has the advantages of short response time, low cost, less packet loss information, and strong privacy protection capabilities. In Reference [21], they proposed a novel dynamic path privacy protection scheme for continuous query service in road networks. This scheme also conceals DPP (dynamic path privacy) users' identities from adversaries; this is provided in the initiator untraceability property of the scheme. The security analysis shows that the model can effectively protect the user's identity anonymously, location information, and service content in LBS. In Reference [22], a fully homomorphic encryption method is used to ensure the safety of the anonymous server itself, and they designed a LBS privacy protection model; the model uses the onion algorithm and asymmetric encryption methods to protect user information.

k-anonymity requires that the same quasi-identifier must have at least k records; each individual record cannot be distinguished from other $k-1$ individuals for sensitive attributes, so the attackers cannot link the records through the quasi-identifier [23–26]. Zhou et al. [27] proposed a neighbor query algorithm that does not rely on trusted third-party anonymous servers to protect the user's location privacy information GHNNQ (Geohash Nearest Neighbor Querying). Guochao et al. [28], according to the rapid search superiority of Geohash coding, proposed a location-based privacy protection method based on interval regions. This approach first generalizes the user's real location to the interval region, then indexes the location with the same code based on the Geohash coding principle as a candidate location set, and then provides personalized k-anonymity privacy protection service for the user, which satisfies the user's privacy requirements.

The filling of the Geohash code can be realized by the Peano curve. Although this curve can fill the space, it has the shortcoming of strong mutability. The common method to solve this problem is to fill the space with the Hilbert curve [29]. In the environment of the road network, using the Voronoi graph [30–32] to partition the road network can solve the defect problem of Geohash Base32, and the Voronoi graph is directly used to prune around the Geohash code region and delete users who are not in the Voronoi unit. On the server side, the user-generated Geohash code is inserted into the prefix tree. The advantages of the prefix tree are as follows: use the common prefix of the string to reduce the query time, minimize the unnecessary string comparison, and the query efficiency is higher than the hash tree. In this paper, we propose a location privacy protection method for road networks based on k-anonymity, in which the Geohash coding model and Voronoi graph are used as grid division principles. Theoretical analysis and experimental results show that it not only achieves privacy protection but also improves the quality of service and data availability.

Comparing the work in this paper with References [27–29], the results are shown in Table 1. Reference [27] proposes the GHNNQ algorithm, which does not rely on third-party servers; that is, the client sends the Geohash-encoded user's location data to the LBS server. By configuring the corresponding query processing algorithm on the server side, the direct interaction between the user and the LBS location server is realized. But it did not combine the actual situation of the road network. Reference [28] uses the superiority of Geohash coding to quickly retrieve information and proposes a location privacy protection method based on interval regions. The user's real location is generalized to the interval area, the same coded location is retrieved as a candidate location set according to the Geohash coding principle, and then according to the user's privacy needs, the user is provided with a personalized k-anonymity privacy protection service. It does not consider the actual situation of the road network and relies on a trusted third-party server. Once the trusted third party is unreliable, it will cause the disclosure of users' privacy. Reference [29] proposed a location privacy protection scheme based on the Hilbert curve. Firstly, a Hilbert curve corresponding to the Hilbert coordinate is generated according to the given coordinate transformation parameters. Secondly, the user's points are transmitted to the location service provider (LSP) through the randomly generated points of the fog server, instead of using k-anonymity and other methods to meet the needs of the location service provider. Then, the weighted KNN algorithm of LSP is used to get the user's interest points. Finally, the POI of the user is transmitted back to the client. This scheme does not use k-anonymity technology and avoids background knowledge attacks and homogeneous attacks. However, this scheme does not combine the actual situation of the road network, nor does it reduce the dimension of the location coordinates. Comparison of related works is shown in Table 1.

3. Preliminaries

3.1. Definitions

Definition 1 (Query information). Generally, the user's query information has the form $Q = (\mathrm{QID}, \mathrm{CGh}, k, t, l_{\min}, R)$. QID represents the user's quasi-identifier, CGh represents the Geohash code, k represents the degree of privacy protection

Table 1: Comparison of related works.

References	Actual road situation considering	Third-party needing	Geohash encoding	k-anonymity technology
Reference [27]	×	×	√	×
Reference [28]	×	√	√	√
Reference [29]	×	×	×	×
This article	√	√	√	√

of the user, t represents the time of sending the request, $l_{\min}$ represents the shortest prefix length of the same code that the user can accept, and R represents the content of the user's query.

Definition 2 (Request information). Generally, request information sent by the server has the form $\mathrm{REQ} = (\mathrm{AS}, R)$. AS represents the formed anonymous set, and R represents the content of the user's query.

Definition 3 (Trie tree). The trie tree is a prefix tree obtained by transforming hash trees that are often used to count and sort large numbers of strings [33]. The idea of a trie tree is to exchange space for time and use the common prefix of strings to reduce the cost of query time to achieve the purpose of improving efficiency.

3.2. Geohash Code. The Geohash code is a geocoding that is essentially used to encode the latitude and longitude into a one-dimensional string. The idea of the Geohash code is to treat the earth as a plane and then divide the longitude and latitude into alternating dichotomies, using binary 0 or 1 to represent the regions divided (the latitude range is -90~90, and the longitude range is -180~180). Every 5 times of division is regarded as a level, and the binary code of each level is converted into a 32-base code (as shown in Table 2), which finally becomes a unique identifier to represent each coordinate on the earth, making it have the characteristics of global uniqueness, multilevel recursion, and one-dimensionality. Since the Geohash code is formed using a dichotomy, it represents a rectangular area rather than a point. The accuracy of the Geohash string is determined by the length of the string, with longer strings having higher accuracy and shorter strings having lower accuracy. When the precision is high enough, the more the prefixes overlap, the closer the two places are, so the Geohash code is often used to query the nearest neighbor.

3.3. Voronoi Diagram. The Voronoi diagram, also known as the Tyson polygon, is a group of continuous polygons composed of vertical bisectors connecting two adjacent points. N points which are different in the plane are divided into planes according to the nearest neighbor principle, and each point is associated with its nearest neighbor region. In this paper, the Voronoi diagram is applied to the road network, and the road network is divided into Voronoi diagram units. The steps to generate the road network Voronoi diagram are as follows. Firstly, the road network is abstracted into a road network model. The undirected graph $G = (V, E)$ is used to represent the road network model, where V represents the intersection point of the road network and E represents the road section between the intersection points. Secondly, based on the road network model, the vertical bisector is drawn for the E, and it is extended to intersect with other vertical lines; then, the polygon formed by these vertical lines is the Voronoi unit. Finally, the above steps are repeated to obtain the Voronoi diagram of the road network intersection. The corresponding Voronoi diagram of the road network is shown in Figure 1.

3.4. LBS Framework Based on IoT. According to the characteristics of information perception and interaction of the IoT and the requirements of location-aware service, the LBS service framework based on the IoT consists of the perception layer, the network layer, the platform layer, and the application layer, as shown in Figure 2.

The perception layer is the foundation of LBS service, which mainly realizes the acquisition of location information, scene information perception, the perception of location-related dynamic spatiotemporal information, etc., and uploads the collected perception information and location data to the network layer.

The network layer mainly realizes the fast, safe, and reliable transmission and exchange of data and transits the perceived data and location information to the platform layer for data sharing and processing. Then, the user's location information and location service request are transmitted to the platform layer, and the intelligent information processing results of the platform layer are returned to the application layer.

The platform layer is based on the cloud computing platform and big data platform to realize the storage and management of massive data. It uses cloud computing, big data, data mining, artificial intelligence, and other technologies to provide personalized location services for the users of the application layer. It also realizes the tracking, control, and monitoring of physical objects in the perception layer.

The application layer is mainly composed of various LBS applications. Users send location information and location service request commands containing service requirements to the LBS server, and the LBS server provides users with location-based navigation, location social interaction, object monitoring, and other responsive interactive services according to the intelligent information processing results of user requests. The LBS system can also use sensors, positioning devices, RFID tags, and other real-time access to the user's current location or activity trajectory and then provide location-related services to the user or automatically achieve the tracking and monitoring of the target object.

Table 2: The Base32.

Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Base32	0	1	2	3	4	5	6	7	8	9	b	c	d	e	f	g
Decimal	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Base32	h	g	k	m	n	p	q	r	s	t	u	v	w	x	y	z

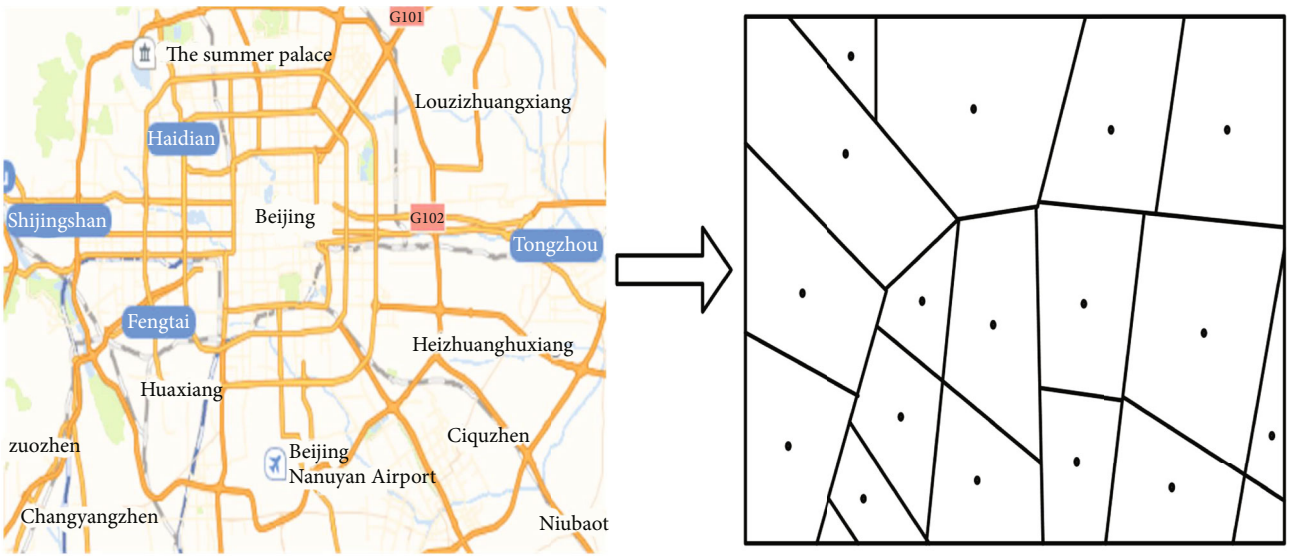


Figure 1: The Voronoi diagram corresponding to the intersection of the road network.

Data transmitted and exchanged between layers of the LBS service framework usually includes personal privacy such as the user's real identity, property account, interests and hobbies, and activity track, as well as business privacy such as the enterprise marketing plan and new product development plan.

From the perspective of the overall architecture of the Internet of Things, location awareness is an indispensable part of the perception layer, which provides the basic location information for the whole Internet of Things system. From the perspective of the application, location-based services will penetrate into many Internet of Things application scenarios to provide differentiated services. Location service is the infrastructure service of the Internet of Things. The intelligent terminal positioning device collects location information, and the location information is provided to the cloud control center as important data. The cloud platform uses the location information of multiple devices to draw a visual interface, which is helpful for the comprehensive analysis and intelligent decision-making of the Internet of Things system.

3.5. System Architecture. The location privacy protection system adopted in this paper consists of three parts: the mobile users, the third-party servers, and the location servers. The system framework is shown in Figure 3.

The system structure of this article is shown in Figure 3, which is mainly composed of a client module, a third-party server, and a location service provider module. The client module is composed of two parts: the positioning module and the conversion mechanism. The positioning module mainly obtains the user's position information through GPS and other positioning devices, and the conversion mechanism converts the user's position coordinates into Geohash codes. The third-party server consists of a database, an anonymous module, and a filtering module. The database is used to store data such as geographic information, road network information, and Geohash codes. The anonymous module selects an anonymous area, generates an anonymous set, and feeds back the anonymity to the database. The location service provider can respond to the query request of the anonymous module and feed back the query result to the screening module, and the screening module will feed back the screening result to the user after screening.

3.6. A Practical Application Scenario. A practical scenario illustrated in Figure 4 is the social network application in smartphones, which brings people a lot of convenience in life. Our intention is to protect and process location information of social applications in smartphones. The client refers to APPs in mobile phones in the social network, and the server refers to location service providers. The server provides the APPs' API interface to obtain the relevant data, adopt the minimum distance grouping algorithm to protect the data, upload the processed data to the APPs' database in the privacy protection processor, adopt the minimum selectivity priority algorithm to achieve secondary protection of the

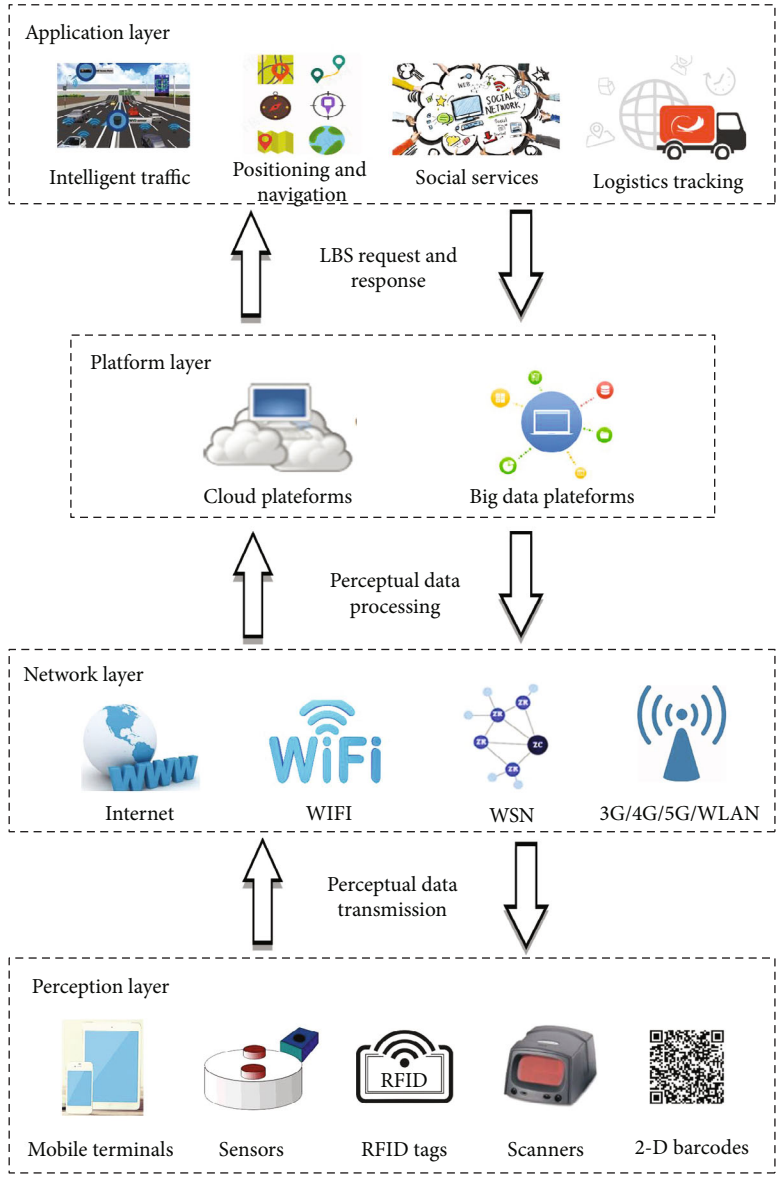


Figure 2: LBS framework based on IoT.

location data, and finally upload the processed data to the database. Users can obtain location query results from the APPs' service providers.

3.7. Attack Model. Almost all LBS providers collect users' personal data, such as identity, location, and interests. Many LBS providers provide different security guarantees, such as Google, Twitter, and YouTube. Once these LBS providers are attacked, users' privacy information will be leaked. The threat model of this paper is shown in Figure 5. The users' location data is acquired through smart mobile devices equipped with positioning technology, such as mobile phones, portable computers, and cars, and the obtained location data is uploaded to the database. Then, the location data is transferred to the LBS servers for further intelligent data processing, which allows users to get convenient services from the LBS providers. The attackers can obtain the user's personal data by attacking the user's smart terminals, LBS servers, or location service providers, which will result in the users' privacy being breached.

4. Algorithm Implementation

In this paper, a location privacy protection algorithm is proposed based on Geohash coding and the Voronoi diagram. The privacy protection algorithm mainly consists of the conversion mechanism and anonymity process. Firstly, the Geohash code generation algorithm is proposed in the conversion mechanism, in which two-dimensional coordinates are transformed into one-dimensional Geohash codes in the mobile client. Then, the G-V anonymity algorithm is proposed in the anonymity process, in which the Voronoi diagram unit is used to protect the location privacy anonymously on the third-party server.

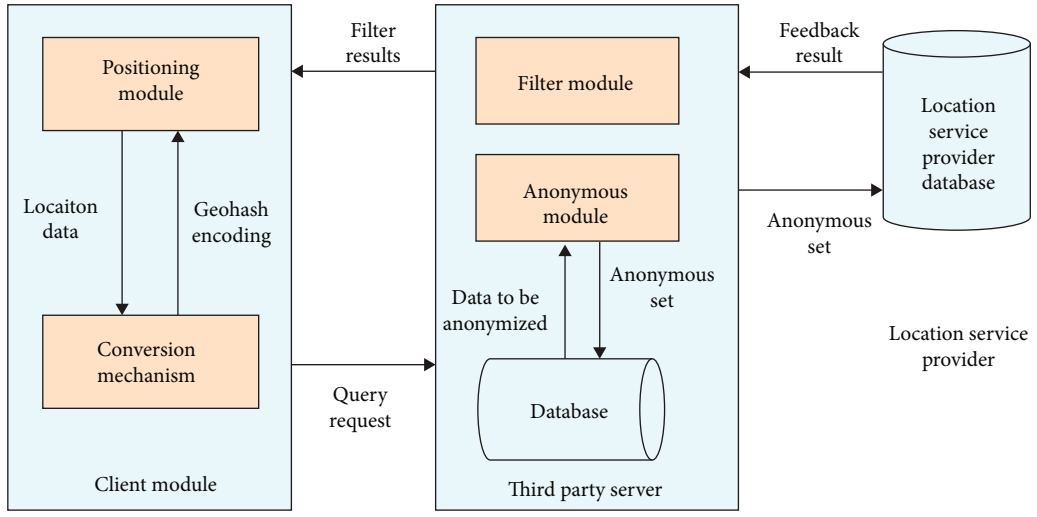


Figure 3: The system architecture.

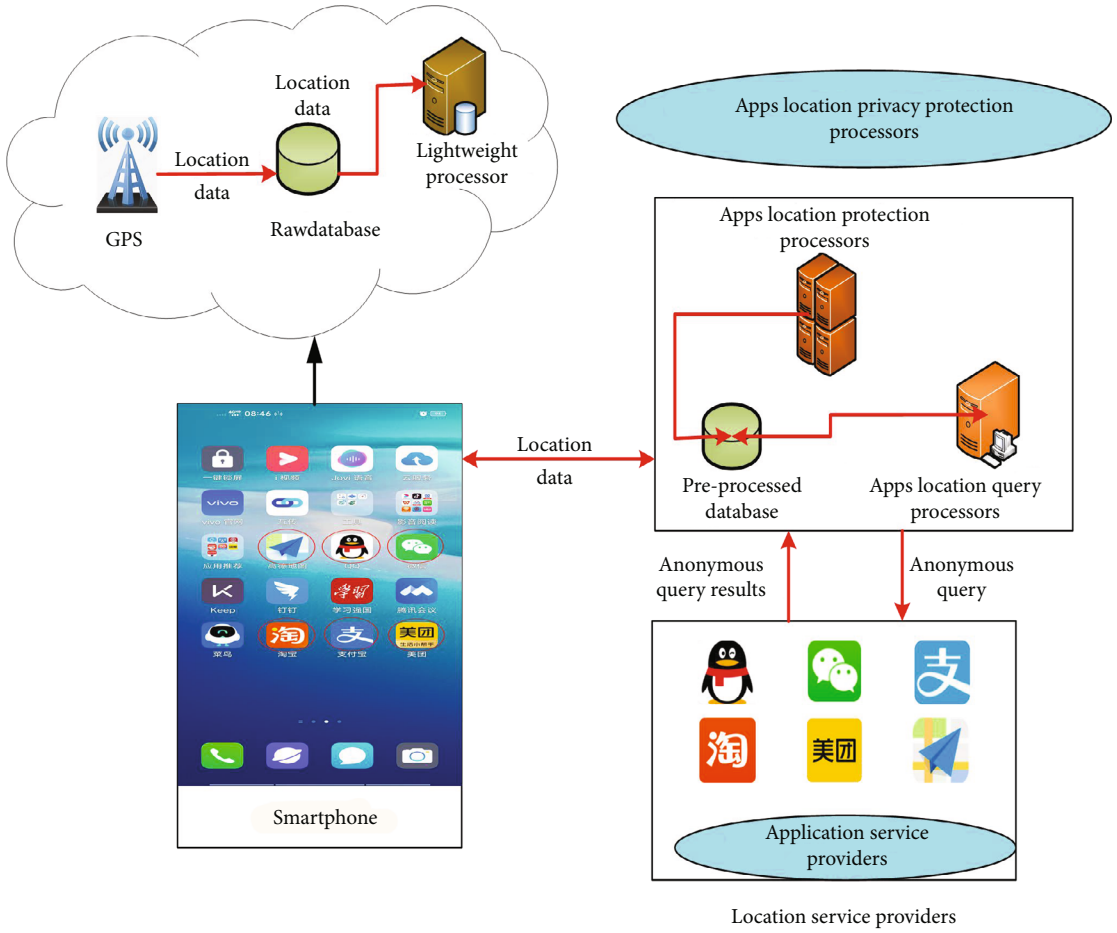


Figure 4: A practical application scenario in smartphones.

4.1. The Conversion Mechanism. The function of the conversion mechanism is to convert the user's location coordinates into codes. Based on the nature of Geohash encoding, the users can choose the degree of accuracy and anonymity. In the Geohash code generation algorithm, the user's latitude and longitude coordinates are bisected and converted into binary firstly, then recording the number of bisection as i. Then, determine the latitude range of the x value of the user

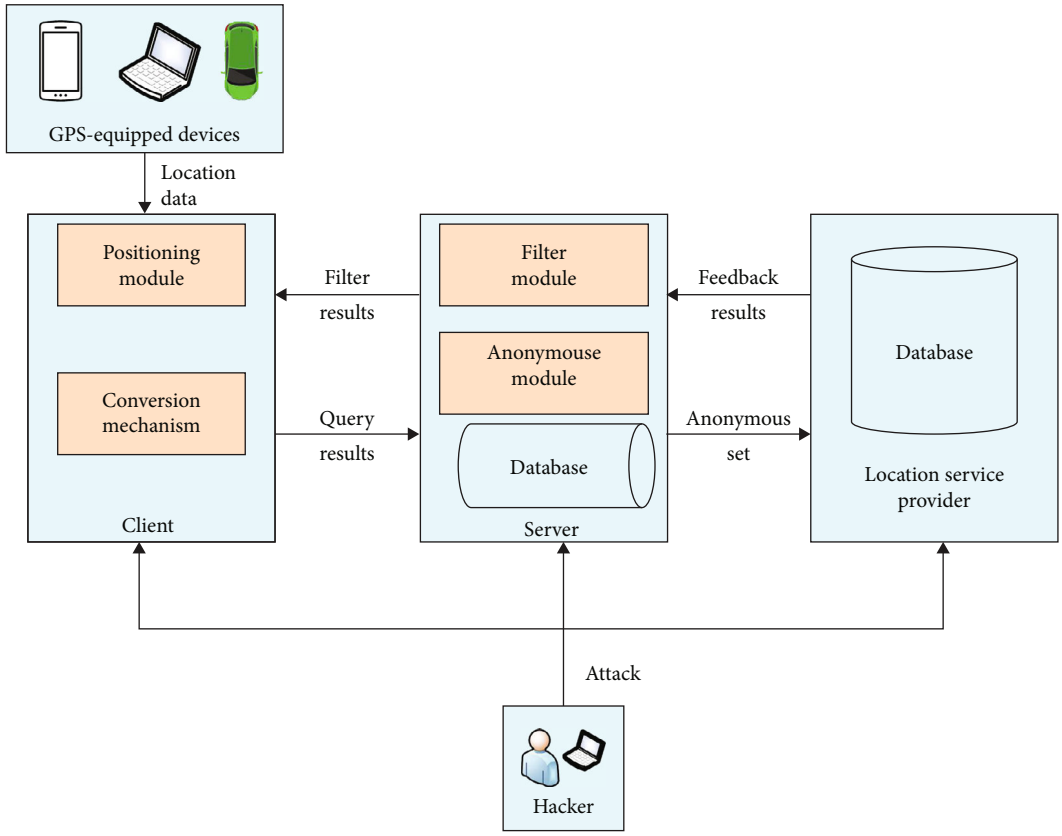


Figure 5: The attack model.

coordinate, and determine the longitude range of the y value of the user coordinate. The left side of the median is marked as 0, and the right side of the median is marked as 1. Finally, Base32 is used for encoding to divide the binary one-dimensional array into a group of five, mapping each group according to Table 2 until the Geohash code output is completed. The Geohash code generation algorithm is as follows.

Take Beijing Tiananmen Square as an example (39.9096 N, 116.3972 E).

As shown in Table 3, the binary code obtained by latitude division is 1011100011. Similarly, the binary code of longitude 116.3972 E is 1101001011 by dividing the longitude region of the earth. The string obtained by combining longitude and latitude is 11100111010010001111. The even bits put the longitude, and the odd bits put the latitude. Every five digits were divided into a group. There are four groups: 11100 (28), 11101 (29), 00100 (4), and 01111 (15). The Geohash code is wx4g, as shown in Table 2.

4.2. G-V Anonymity Algorithm. In this section, the G-V anonymity algorithm is proposed based on the Geohash coding model and Voronoi diagram meshing principle. The basic idea of the algorithm is to find k nearest neighbors in the anonymous region, combine the Geohash with the Voronoi diagram, determine the nearest neighbor by using the characteristics of the Geohash code, delete the mutative users by using the Voronoi diagram, and finally realize the anonymity protection of location privacy. The details are as follows. Firstly, the user who makes the request at time t forms a set to be anonymous in the database and uploads the relevant data to the anonymous module. Secondly, generate a prefix tree to initialize the root node, insert the Geohash code of the unknown user into the prefix tree, and traverse the Geohash string. If there is such a character in the prefix tree, add one to the original number; otherwise, allocate a new node and record the number of new characters until all characters are inserted. Then, query the strings with the same prefix. If there is no application for generating dummy elements, the number of users will be counted, and the anonymous users with the same prefix but not in the Voronoi diagram unit will be deleted to form a new array $U*$. Finally, judge whether the number of users in the new array meets the$k$value; if it is satisfied, output it directly; otherwise, the corresponding number of dummy$D = k - m$will be generated. Dummy location technology is also a fake location technology, and k -anonymity can also be achieved by adding fake locations. The dummy location technology requires that in the query process, in addition to the real location, a number of additional fake location information must be added. The server not only responds to requests for real locations but also responds to requests for fake locations so that the attacker cannot tell which is the real location of the user.

The G-V anonymity algorithm is as follows.

4.3. The Algorithm Analysis

4.3.1. Security Analysis. The client converts the user's location coordinates into one-dimensional Geohash codes.

Input: Position coordinates (x, y), code length l
Output: Geohash code (CGh)
1. lat = $\{-90.0, 90.0\}$, lon = $\{-180.0, 180.0\}$;
2. numbits = $(l * 5)/2$; //The separate code length of longitude and latitude
3. Base32 = {'0,' '1,' '2,' '3,' '4,' '5,' '6,' '7,' '8,' '9,' 'b,' 'c,' 'd,' 'e,' 'f,' 'g,' 'h,' 'j,' 'k,' 'm,' 'n,' 'p,' 'q,' 'r,' 's,' 't,' 'u,' 'v,' 'w,' 'x,' 'y,' 'z'};
4. int $i = 0$, $j = 0$; //i is the number of latitudinal dichotomies, and j is the number of longitude dichotomies
5. latmid = lat/2, lonmid = lon/2; //Dichotomizing the latitude and the longitude
6. while (latmid) //Convert longitude coordinates to binary
7. {
8. char lat$B[i] = 0$;
9. if (x> = latmid)
10. lat$B[i] = 1$;
11. else lat$B[i] = 0$;
12. $i++$;
13. }
14. while (lonmid) //Convert latitude coordinates to binary
15. {
16. char lon$B[j] = 0$;
17. if (y> = lonmid)
18. lon$B[j] = 1$;
19. else lon$B[j] = 0$;
20. $j++$;
21. }
22. while (lat$B[i]$! = Null || lon$B[j]$! = Null) //The longitude and latitude are combined, and the even bit is used to put the longitude, and the odd bit is used to put the latitude
23. {
24. int $k = 0$;
25. char $B[k]$;
26. if (k%2 == 0)
27. {
28. $B[k]$ = lon$B[j]$;
29. $j++$;
30. $k++$;
31. }
32. else
33. {
34. $B[k]$ = lat$B[i]$;
35. $i++$;
36. $k++$;
37. }
38. return $B[k]$;
39. }
40. int $m = 0$;
41. for ($k = 0$; $k +$ = 4)
42. {
43. $B[m] = B[k, k + 4]$; //Divide $B[k]$ into groups of five digits
44. $B[m] \longrightarrow$ Base32; //Map binary to Base32
45. $m++$;
46. CGh = $B[m]$;
47. }
48. return CGh; //Get Geohash code

Algorithm 1: Geohash code generation algorithm.

Geohash uses a string to represent the two coordinates of longitude and latitude. It represents a rectangular area. Every point in the rectangular area may be the exact location of the user, which makes the attacker unable to determine the exact location of the user. The longer the Geohash-encoded string is, the smaller the rectangular area is, the more accurate the result is, but the weaker the privacy protection is. Because the Geohash string encoding algorithm is implemented in the client, the privacy protection strength can be controlled by the user. This is the first layer of protection for location privacy.

In the third-party server, according to the user's anonymity requirement, we can find no less than k nearest neighbors in the Voronoi unit to form an anonymous set. The

Table 3: The calculation of latitude.

Latitude region	Left interval 0	Right interval 1	39.909
(-90.0, 90.0)	(-90.0, 0.0)	(0.0, 90.0)	1
(0.0, 90.0)	(0.0, 45.0)	(45.0, 90.0)	0
(0.0, 45.0)	(0.0, 22.5)	(22.5, 45.0)	1
(22.5, 45.0)	(22.5, 33.75)	(33.75, 45.0)	1
(33.75, 45.0)	(33.75, 39.37)	(39.38, 45.0)	1
(39.375, 45.0)	(39.375, 42.18)	(42.19, 45.0)	0
(39.375, 42.19)	(39.375, 40.78)	(40.78, 42.19)	0
(39.375, 40.78)	(39.375, 40.07)	(40.08, 40.78)	0
(39.375, 40.08)	(39.375, 39.72)	(39.73, 40.08)	1
(39.73, 40.08)	(39.73, 39.90)	(39.90, 40.08)	1

maximum probability that an attacker can lock a user is $p(ui) = 1/k$. The higher the value of k, the lower the probability of an attacker finding the user. Moreover, in the anonymous area formed by the Voronoi unit, the user's specific location is further blurred. k-anonymity technology can guarantee the following three points. (1) The attacker cannot know whether a specific individual is in the public data. (2) Given a person, the attacker cannot confirm whether he has a certain sensitive attribute. (3) The attacker cannot confirm which person a piece of data corresponds to. This is the second layer of protection for location privacy.

To sum up, even if the attacker intercepts the information sent by the user, he cannot find the specific location of the user.

4.3.2. Complexity Analysis. There is no need to use the Euclidean distance calculation when finding the k nearest neighbors; just find the user with the same prefix who sent the request at time t. In this paper, a *trie* tree is used to search. The time complexity of finding the nearest neighbor is related to the length of the Geohash code, so its time complexity is linear $O(l)$, where l is the length of the code. The disadvantage of the *trie* tree is to trade space for time. Base32 used in this paper has 32 characters, so the space complexity is $O(32^n)$. When there are few users and when there is a need to generate dummy elements, the time complexity of successfully constructing an anonymous set depends on the number of generated dummy $D = k - N$, which is linear order $O(D)$.

4.3.3. Mathematical Analysis. The central idea of the algorithm in this paper is as follows. Firstly, convert the user's location coordinates into a Geohash code of the corresponding length on the client side. Secondly, insert the Geohash code into the prefix tree on the server side, and the prefix tree finds the nearest neighbors based on the characteristics of the coded similar prefixes, and the Voronoi diagram divides the area unit to complete the pruning. Finally, according to the Geohash coding model and the Voronoi diagram grid division principle, the G-V anonymity algorithm is proposed. The basic idea of the algorithm is to find k nearest neighbors in the anonymous area to meet the k -anonymity requirement.

According to the G-V anonymity algorithm selection scheme, the anonymous set obtained is $c, c = \{l_1, l_2, \cdots, l_{k-1}, l_k\}$, that is, $|c| = k$. Then, the probability Q that the attacker obtains the user's real location from the anonymous set is calculated as follows:

$$Q = \frac{1}{|c|} = \frac{1}{k}. \tag{1}$$

The k-anonymity technology can ensure that each individual record stored in the release dataset cannot be distinguished from other $k-1$ individuals for sensitive attributes; that is, the k-anonymity mechanism requires at least k records of the same quasi-identifier. The implementation of k-anonymity technology makes it impossible for observers to identify users through quasi-identifiers with a confidence higher than $1/k$, so k-anonymity technology is effective in terms of privacy protection.

5. Experimental Analysis

In this section, we analyze the experimental results of the G-V anonymous location privacy protection method and illustrate the actual performance of the method by analyzing the results of four indicators: anonymous success rate, privacy protection degree, algorithm running time, and algorithm antiattack on simulated datasets.

The algorithm is compared with the G-Casper and Casper algorithms. The G-Casper algorithm uses a bottom-up mechanism to make a string fuzzy query on the Geohash code of the target location to determine the $k-1$ nearest neighbors of the anonymous region. When expanding the scanning area, it scans the grid of the user and the surrounding grid across the domain and then performs hierarchical recursion. At the same time, it uses two parameters $L_{\max}$ and $L_{\min}$ to control the anonymous region. Finally, the redundant grid is removed by the pruning algorithm, and a candidate grid area is randomly sent to replace the user's original location to achieve the effect of k-anonymity. The idea of the Casper algorithm is to adopt the quadtree architecture, select the vertical grid and horizontal grid which meet the conditions each time, and store the information of each grid in the form of quadtree layer by layer until the grid is reclassified into a region and satisfies k-anonymity and minimum anonymity.

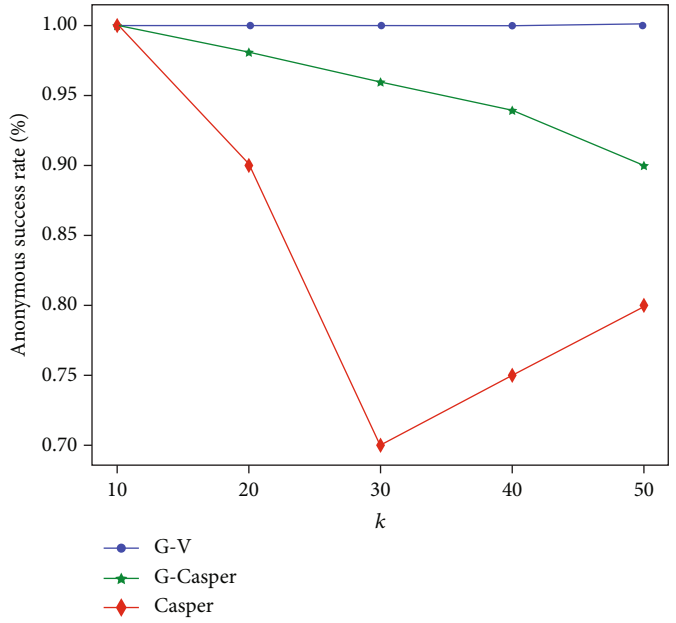
5.1. Environment Setting. The hardware environment is as follows: Intel® Core i5, CPU 1.7 GHz, and memory 4.00 GB. The software environment is as follows: Windows 10 64-bit operating system, compiled language using the C++ language. The experiment is implemented using Python. The experiment is carried out on two datasets. One is the POI dataset [34] from several provinces in China, which contains about 4 million pieces and is stored in the MySQL database. One is the Gowalla dataset [35]. The algorithm proposed in this paper is compared with G-Casper [36] and Casper [37] algorithms in terms of the anonymous success rate, privacy protection degree, algorithm running time, and algorithm antiattack.

Input: The user set to be anonymous $U = \{u_1, u_2, u_3, \cdots, u_n\}$, the Voronoi diagram of the road network, k, $l_{\min}$
Output: The anonymous set AS
1. send($Q\{u_1, u_2, u_3, \cdots, u_n\}$); //Sending anonymous request from users
2. receive($Q\{u_1, u_2, u_3, \cdots, u_n\}$); //The server received information from n users
3. send($Q\{u_1, u_2, u_3, \cdots, u_n\}$ Voronoi); //Sending data to the anonymous module
4. receive($Q\{u_1, u_2, u_3, \cdots, u_n\}$ Voronoi); //The anonymous module accepts users' data
5. int Max = 32; //Max represents the size of the character set
6. int count = 0; //count represents the number of times the character appears
7. struct trieNode next[Max] //The next array represents the type of each character
8. for ($i = 0$; $i <$ Max ; i++) //Building the prefix tree
9. {
10. if ($i = 0$)
11. next[i] = Null; //Initializing root node
12. else
13. next[i] = CGh; //Insert the Geohash code into the prefix tree
14. }
15. while (prefix tree! = Null) //Traverse the prefix tree
16. {
17. if (next[i] == CGh)
18. {
19. count + +;
20. i + +;
21. }
22. if (next[i]! = CGh)
23. {
24. next[i] = newnode[++num]; //The node is none. Assigning a new node
25. count + +;
26. }
27. }
28. while (search! = Null) //Querying strings with the same prefix
29. {
30. if (search'CGh' == Null)
31. return 0;
32. else
33. {search + +;
34. return m = search; //Number of users with the same prefix
35. }
36. }
37. $U* = \{u_1, u_2, u_3, \cdots, u_m\}$
38. if ($m >= k$)
39. AS $\longleftarrow U*$;
40. else
41. {
42. $D = k - m$; //Generating dummy D
43. AS $\longleftarrow U* + D$;
44. }
45. return AS; //Anonymous success

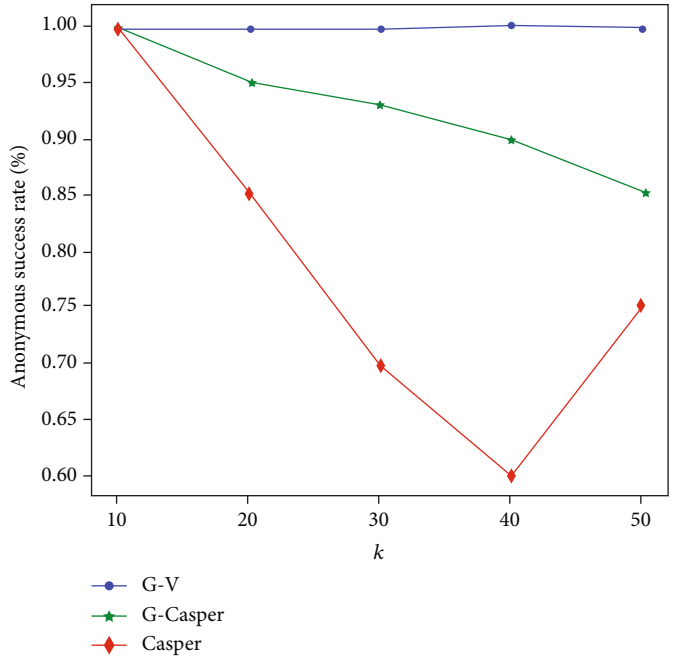
Algorithm 2: The G-V anonymity algorithm.

5.2. Anonymous Success Rate. When the Geohash code length l is unchanged, the algorithm proposed in this paper is compared with the G-Casper and Casper algorithms in terms of the anonymous success rate, and the result is shown in Figure 6. The x-axis represents the size of the anonymous area, and the y-axis represents the anonymous success rate. The anonymous success rate of the G-V anonymity algorithm has nothing to do with the value of k, while the anonymous success rate of the G-Casper and Casper algorithms decreases with the increase of the value of k. Therefore, compared with the G-Casper and Casper algorithms, the algorithm proposed in this paper has the best anonymous success rate, followed by the G-Casper algorithm. As the value of k increases, the anonymous area of the Casper algorithm is too large to be accurate enough, and it requires a lot of storage space to store the information of each grid. Therefore, the Casper algorithm has the worst anonymous success rate.

When the anonymous region k is unchanged, the algorithm proposed in this paper is compared with the G-Casper and Casper algorithms in terms of the anonymous success rate. The result is shown in Figure 7. The x-axis



(a) Gowalla

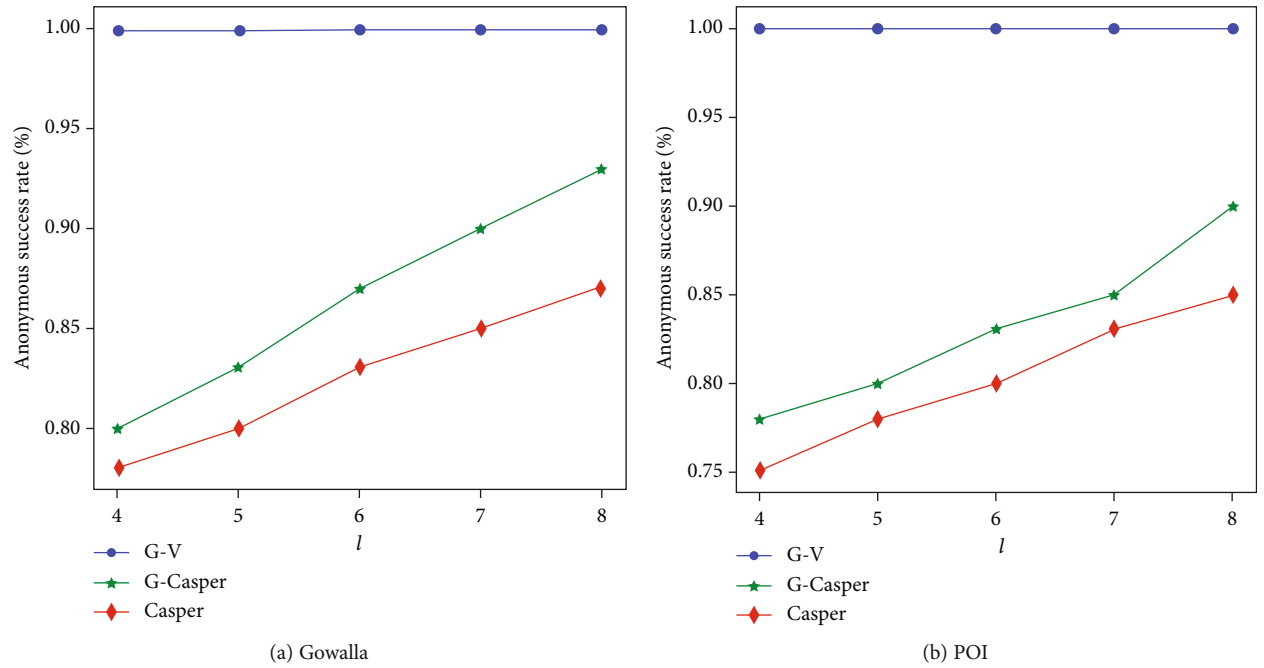


(b) POI

Figure 6: The anonymous success rate of the algorithm when the l is unchanged.

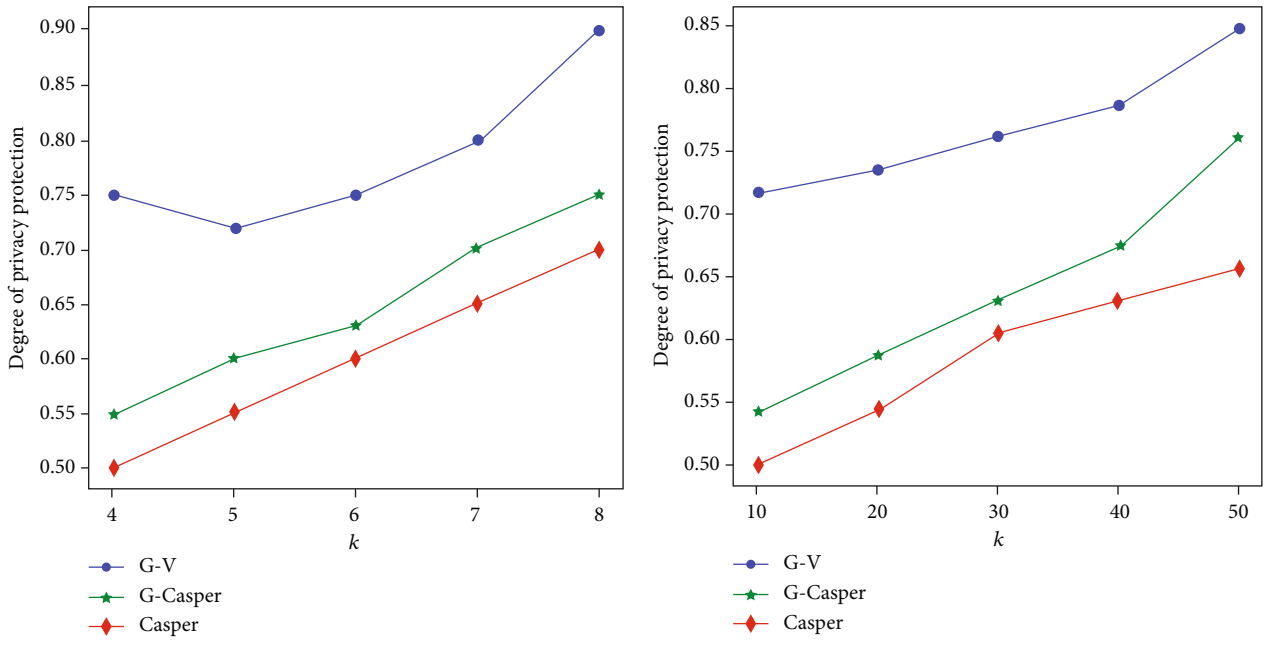
represents the length of the Geohash code, and the y-axis represents the anonymous success rate. The anonymous success rate of the G-V anonymity algorithm has nothing to do with the value of l, while the anonymous success rate of the G-Casper and Casper algorithms decreases with the increase of the value of l. Therefore, compared with the G-Casper and Casper algorithms, the algorithm proposed in this paper has the best anonymous success rate, followed by the G-Casper algorithm, and the Casper algorithm has the worst anonymous success rate.

5.3. Degree of Privacy Protection. The proposed algorithm is compared with the G-Casper and Casper algorithms in terms of privacy protection under the condition that the Geohash encoding length l remains unchanged. The results are shown in Figure 8. The x-axis represents the size of the anonymous region, and the y-axis represents the degree of privacy protection. The larger the k value is, the lower the risk of leakage is, and the better the privacy protection is. With the increase of the k value, the degree of privacy protection increases. The privacy protection of the G-V anonymity algorithm is



(a) Gowalla (b) POI

Figure 7: The anonymous success rate of the algorithm when the k is unchanged.

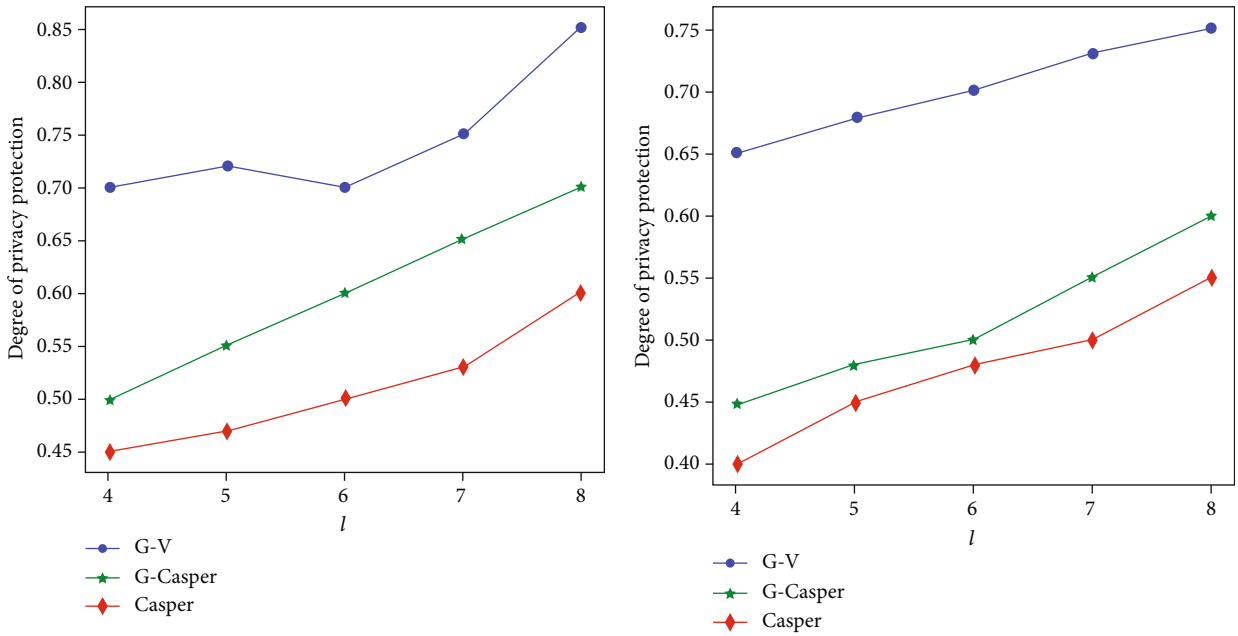


(a) Gowalla (b) POI

Figure 8: The degree of privacy protection of the algorithm when the l is unchanged.

affected by two factors: k and l. When the value of k increases, the number of users in the anonymous area increases, and the probability of the attacker inferring the user's location decreases. Therefore, the privacy protection degree of the algorithm proposed in this paper is the best, followed by the G-Casper algorithm. With the increase of the k value, the anonymous region of the Casper algorithm is too large to be accurate, and the privacy protection degree is poor.

The proposed algorithm is compared with the G-Casper and Casper algorithms in terms of privacy protection under the condition that the anonymous region$k$ remains unchanged. The results are shown in Figure 9. The x-axis represents the length of Geohash encoding, and the y-axis represents the degree of privacy protection. The higher the l value is, the lower the risk of leakage is, and the better the privacy protection is. The G-V anonymity algorithm is affected



(a) Gowalla (b) POI

Figure 9: The degree of privacy protection of the algorithm when the k is unchanged.

by the encoding length l. When the encoding length l increases, the probability of the attacker guessing the user's location decreases. Therefore, the proposed algorithm has the best degree of privacy protection, followed by the G-Casper algorithm, which has a poor degree of privacy protection.

5.4. Algorithm Running Time. When the length of the Geohash code l remains unchanged, the algorithm proposed in this paper is compared with the G-Casper and Casper algorithms in terms of algorithm running time. The results are shown in Figure 10. The x-axis represents the size of the anonymous area, and the y-axis represents the running time of the algorithm. The larger the anonymous area, the more time the algorithm will run. Regardless of the value of k, the anonymity time of the algorithm proposed in this article will not fluctuate much. Therefore, the running time of the algorithm proposed in this paper is less, followed by the G-Casper algorithm, and the Casper algorithm requires more time.

In the case of the anonymous region when$k$ is unchanged, the algorithm in this paper is compared with the G-Casper and Casper algorithms in terms of algorithm running time. The results are shown in Figure 11. The x -axis represents the encoding length, and the y-axis represents the running time of the algorithm. The larger the encoding length l is, the longer the algorithm runs. The algorithm proposed in this paper uses the prefix tree structure, so the query time is less, followed by the G-Casper algorithm, and the Casper algorithm requires more time.

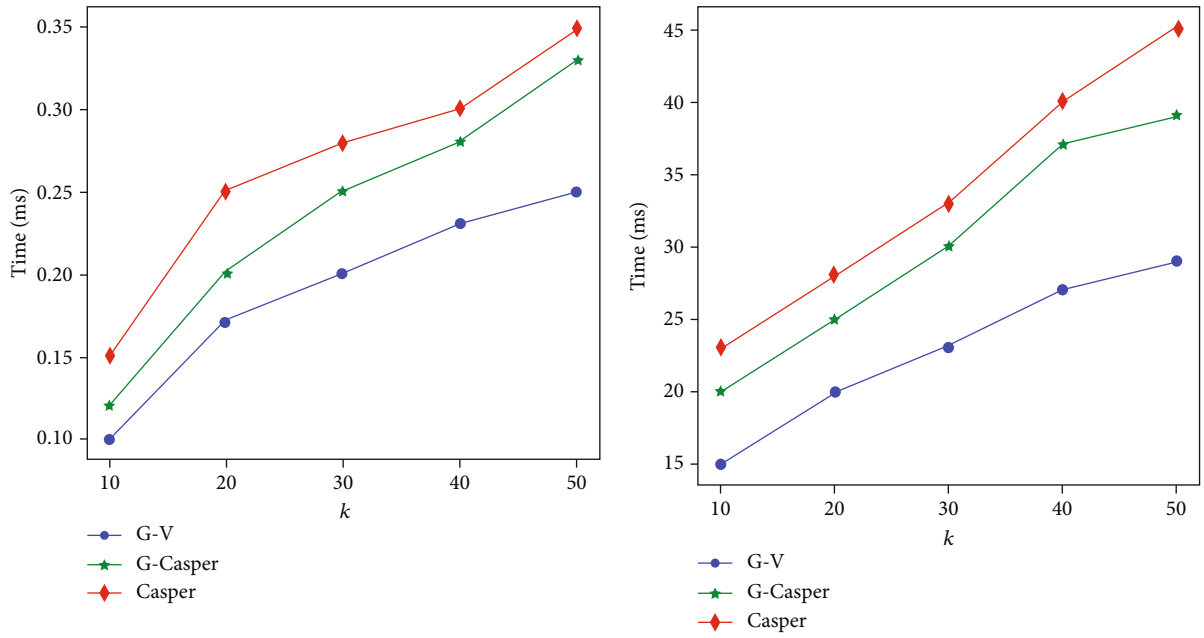
5.5. The Antiattack Ability. The privacy protection object in data publishing is mainly the corresponding relationship between the user's sensitive data and individual identity. Generally, the way of deleting identifiers to publish data cannot really prevent privacy disclosure. Attackers can obtain individual privacy data through link attacks.

The chain attack refers to the link operation between the published data and the external data obtained from other channels to infer the privacy data, thus causing privacy leakage, which is equivalent to an expansion of the dimension of personal information. The simplest example is that two tables in the database get more information through primary key association.

In order to solve the problem of privacy leakage caused by link attacks, the k-anonymity method is introduced. k -anonymity publishes data with low precision through generalization and concealment technology, which makes each record at least have the same quasi-identifier attribute value as other $k-1$ records in the data table, so as to reduce the privacy leakage caused by link attacks.

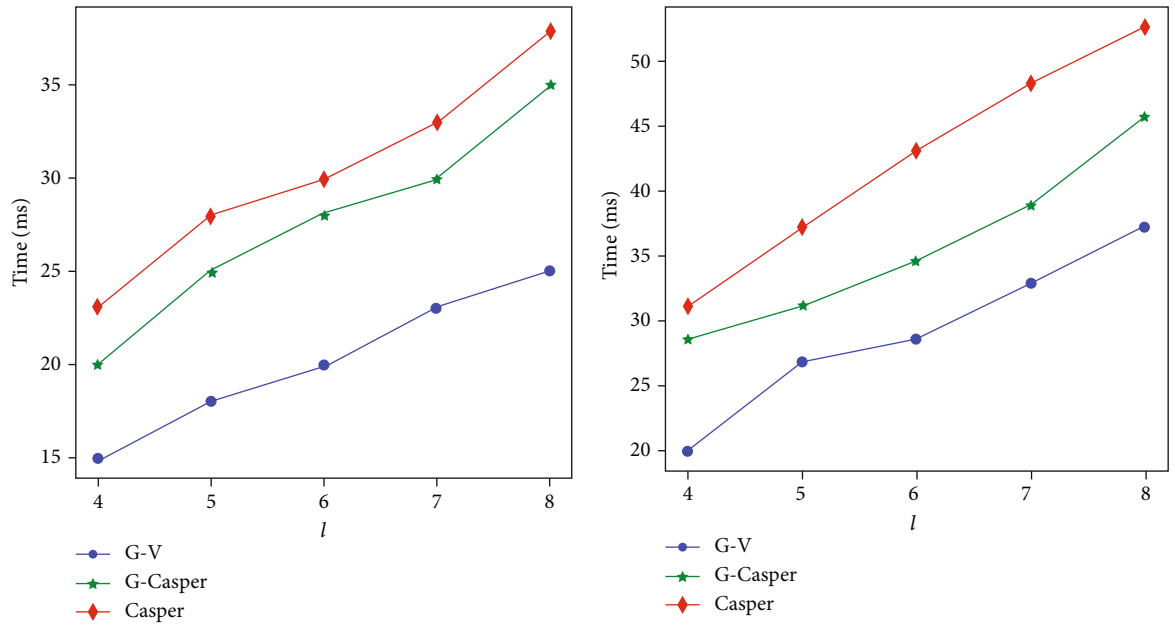
Under the condition that the length l of the Geohash code is unchanged, the algorithm in this paper is compared with the G-Casper and Casper algorithms in terms of antiattack. The result is shown in Figure 12. The x-axis represents the size of the anonymous area, and the y-axis represents the attack resistance of the algorithm. The larger the anonymous area, the stronger the algorithm's resistance to chain attacks. The algorithm proposed in this paper converts the location coordinates into Geohash codes, and Geohash codes represent a region, not a location. Therefore, the location is blurred, showing strong resistance to attacks. The second is the G-Casper algorithm. The Casper algorithm is weaker against attacks.

When the anonymous domain k is unchanged, the algorithm in this paper is compared with the G-Casper and Casper algorithms in terms of antiattack. The result is shown in Figure 13. The x-axis represents the length of the Geohash code, and the y-axis represents the attack resistance of the



(a) Gowalla (b) POI

Figure 10: The running time of the algorithm when the l is unchanged.

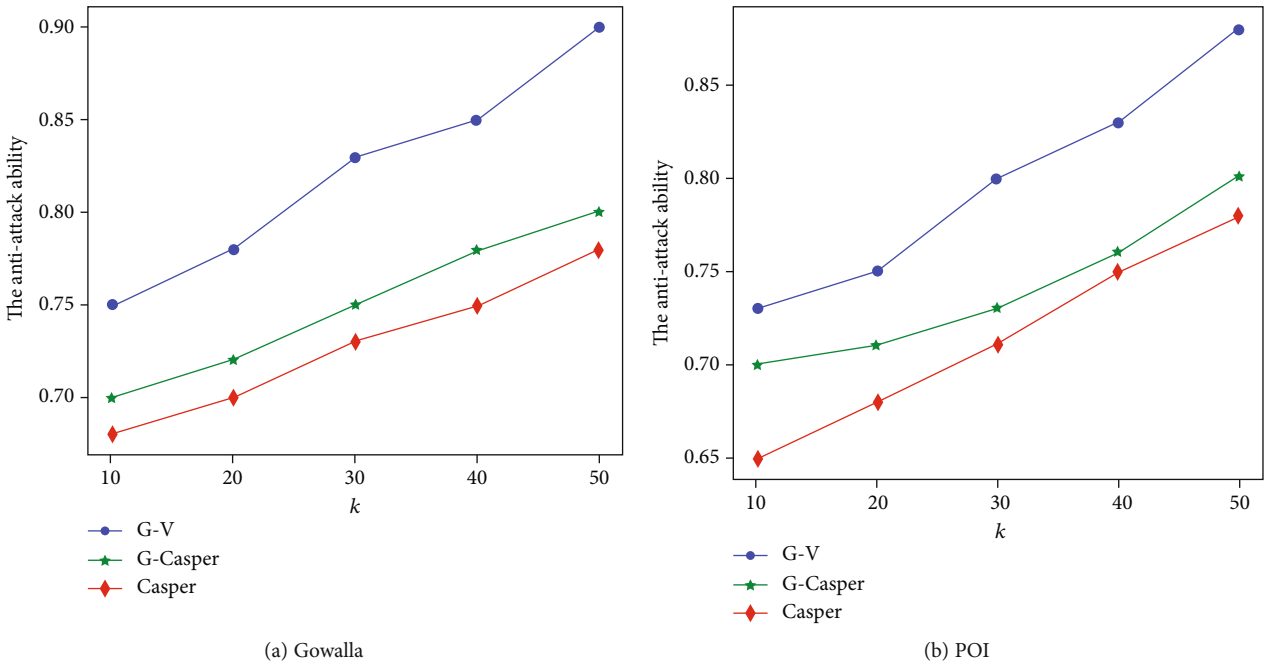


(a) Gowalla (b) POI

Figure 11: The running time of the algorithm when the k is unchanged.

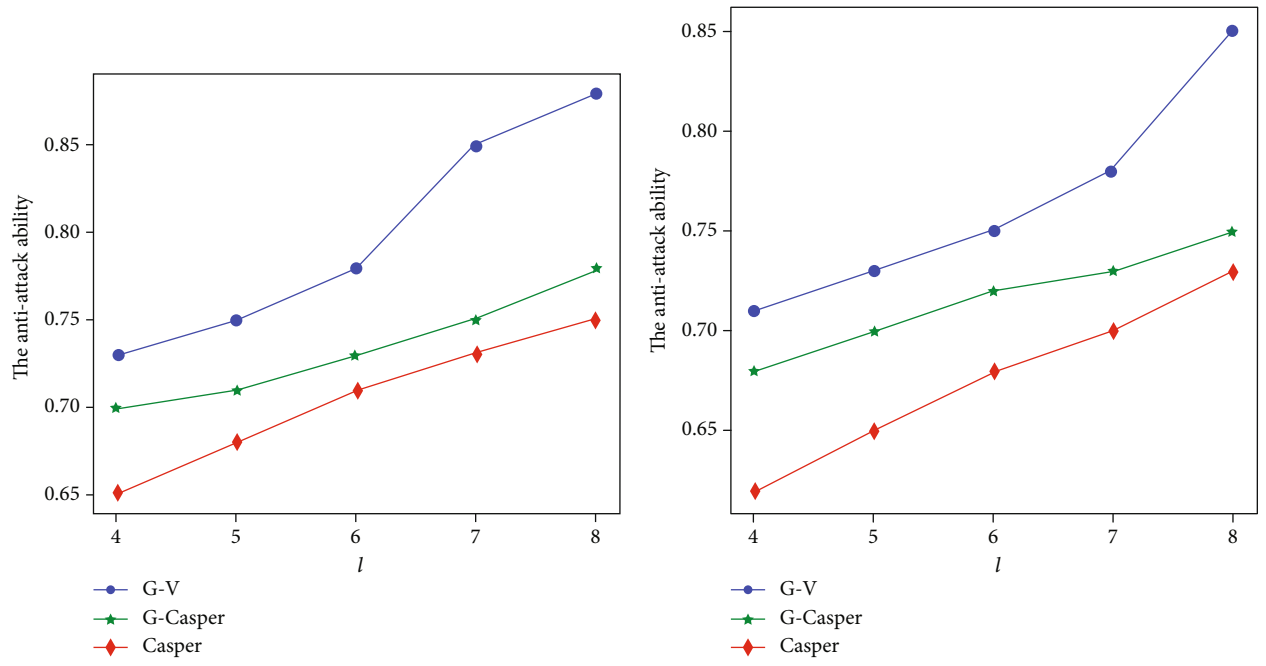
algorithm. The larger the code length, the stronger the algorithm's resistance to chain attacks. The algorithm proposed in this paper adopts a prefix tree structure, which has a strong antiattack, followed by the G-Casper algorithm, and the Casper algorithm is weaker.

5.6. Discussion Section. The design features of the algorithm in this paper are as follows. On the client side, the Geohash algorithm is proposed, which converts the user's location coordinates into a Geohash code of the corresponding length. On the server side, the Geohash code generated by the user is inserted into the prefix tree, the prefix tree is used to find the nearest neighbors according to the characteristics of the coded similar prefixes, and the Voronoi diagram is used to divide the area units to complete the pruning. Then, using the Geohash coding model and the Voronoi diagram grid



(a) Gowalla (b) POI

Figure 12: The antiattack of the algorithm when the l is unchanged.



(a) Gowalla (b) POI

Figure 13: The antiattack of the algorithm when the k is unchanged.

division principle, the G-V anonymity algorithm is proposed to find k neighbors in an anonymous area so that the user's location data meets the k-anonymity requirement in the area unit, thereby achieving anonymity protection of location privacy.

Traditional location privacy protection methods mostly use GPS locations to calculate anonymous areas. Anonymous servers use an anonymous area that satisfies k-anonymity instead of the user's real location for query processing. This often requires the server and the user to perform a lot of GPS calculations. This article uses the Geohash code instead of GPS location and then inserts the Geohash code into the prefix tree, which saves time. Since the G-V anonymity algorithm assumes that the third parties and mobile users involved in the calculation are credible, future research work can be carried out on semitrusted or untrusted third parties

or users in the system. The *k*-anonymity technology can resist chain attacks but cannot resist background knowledge attacks, homogenization attacks, and supplementary data attacks. Future work will consider how to better resist the above attacks and better protect users' privacy.

6. Conclusions

The rapid development of IoT technology promotes the rise of LBS, which brings a lot of convenience to people's lives, but it also brings severe challenges to the privacy security of users. Location data usually relates to the user's privacy information. Once the location information is leaked, it will bring serious threats to the user's privacy. In this paper, based on the Geohash coding model and Voronoi diagram meshing principle, we propose a location privacy protection method for road networks. We use the third-party server architecture. In the third-party server, the prefix tree is used to compare the Geohash code to find the nearest neighbors. Voronoi is used to complete the pruning. Then, the number of remaining users determines whether to generate dummy elements. The dimensional coordinates are reduced to a one-dimensional array, avoiding the calculation of floating-point numbers. The experimental result shows that the proposed anonymity algorithm has the advantages of short anonymity time, high success rate, and good service quality while ensuring privacy protection. Since the G-V anonymity algorithm assumes that the third party and mobile user participating in the operation are trusted, future research work can be carried out on the existence of the semitrusted or untrusted third party or user in the system.

Data Availability

All data, models, and codes generated or used during the study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China under grant nos. 61702316 and 61872088, the Natural Science Foundation of Shanxi Province under grant nos. 201801D221177 and 201901D111280, the Educational Research Projects of Young and Middle-Aged Teachers in Fujian Education Department under grant no. JAT170142, the Key Research and Development Project of Shandong Province under grant no. 2019JZZY010134, the Graduate Education Reform Research Project of Shanxi Province under grant no. 2020YJJG145, the Guangxi Key Laboratory of Trusted Software under grant no. KX202042, the Opening Foundation of Fujian Provincial Key Laboratory of Network Security and Cryptology Research Fund, Fujian Normal University under grant no. NSCL-KF2021-06, and the Natural Science Foundation of Fujian Province under grant no. 2019J01276.

References

[1] H. Huang, G. Gartner, J. M. Krisp, M. Raubal, and N. van de Weghe, "Location based services: ongoing evolution and research agenda," *Journal of Location Based Services*, vol. 12, no. 2, pp. 63–93, 2018.

[2] R. Gupta and U. P. Rao, "A hybrid location privacy solution for mobile LBS," *Mobile Information Systems*, vol. 2017, Article ID 2189646, 11 pages, 2017.

[3] R. Gupta and U. P. Rao, "VIC-PRO: vicinity protection by concealing location coordinates using geometrical transformations in location based services," *Wireless Pers Commun*, vol. 107, no. 2, pp. 1041–1059, 2019.

[4] R. Gupta and U. P. Rao, "Achieving location privacy through CAST in location based services," *Journal of Communications & Networks*, vol. 19, no. 3, pp. 239–249, 2017.

[5] X. S. Xue and J. F. Chen, "Using Compact Evolutionary Tabu Search algorithm for matching sensor ontologies," *Swarm and Evolutionary Computation*, vol. 48, pp. 25–30, 2019.

[6] X. S. Xue, X. Wu, C. Jiang, G. Mao, and H. Zhu, "Integrating sensor ontologies with global and local alignment extractions," *Wireless Communications and Mobile Computing*, vol. 2021, 10 pages, 2021.

[7] J. B. Xiong, J. Ren, L. Chen et al., "Enhancing privacy and availability for data clustering in intelligent electrical service of IoT," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1530–1540, 2019.

[8] J. B. Xiong, R. Ma, L. Chen et al., "A personalized privacy protection framework for mobile crowdsensing in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4231–4241, 2020.

[9] J. Zhang, B. Zhong, B. Fang, and J. Ding, "An improvement of track privacy protection method based on K-anonymity technology," *Intelligent Computer and Applications*, vol. 9, no. 5, 2019.

[10] A. Mille, L. Karim, J. Almhana, and N. Khan, "Location privacy-preserving mobile crowd sensing with anonymous reputation," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 1812–1817, Limassol, Cyprus, 2020.

[11] N. Ravi, C. M. Krishna, and I. Koren, "Enhancing vehicular anonymity in ITS: a new scheme for mix zones and their placement," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10372–10381, 2019.

[12] R. Zhang, X. Wang, P. Cheng, and J. Chen, "A novel pseudonym linking scheme for privacy inference in VANETs," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–5, Antwerp, Belgium, 2020.

[13] Q. A. Arain, D. Zhongliang, I. Memon et al., "Privacy preserving dynamic pseudonym-based multiple mix-zones authentication protocol over road networks," *Wireless Personal Communications*, vol. 95, no. 2, pp. 505–521, 2017.

[14] L. Benarous, B. Kadri, S. Bitam, and A. Mellouk, "Privacy-preserving authentication scheme for on-road on-demand refilling of pseudonym in VANET," *International Journal of Communication Systems*, vol. 33, no. 10, 2020.

[15] X. Ye, J. Zhou, Y. Li, M. Cao, D. Chen, and Z. Qin, "A location privacy protection scheme for convoy driving in autonomous driving era," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1388–1400, 2021.

[16] L. Zhao, X. Wang, and X. Huang, "Verifiable single-server private information retrieval from LWE with binary errors," *Information Sciences*, vol. 546, no. 2, 2021.

[17] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 322–329, 2019.

[18] J. Li, X. Kuang, S. Lin, X. Ma, and Y. Tang, "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes," *Information Sciences*, vol. 526, pp. 166–179, 2020.

[19] S. Gupta and G. Arora, "Use of homomorphic encryption with GPS in location privacy," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 42–45, Mathura, India, 2019.

[20] I. Memon, Q. A. Arain, H. Memon, and F. A. Mangi, "Efficient user based authentication protocol for location based services discovery over road networks," *Wireless Personal Communications*, vol. 95, no. 4, pp. 3713–3732, 2017.

[21] I. Memon and Q. A. Arain, "Dynamic path privacy protection framework for continuous query service over road networks," *World Wide Web*, vol. 20, no. 4, pp. 639–672, 2017.

[22] Q. A. Arain, R. A. Shaikh, and H. Memon, "User privacy protection based on road network model for location based services," *Journal of Information & Communication Technology (JICT)*, vol. 10, 2016.

[23] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, "A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services," *Future Generation Computer Systems*, vol. 94, pp. 40–50, 2019.

[24] K. Zhou and J. Wang, "Trajectory protection scheme based on fog computing and K-anonymity in IoT," in *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 1–6, Matsue, Japan, 2019.

[25] X. Yang, L. Gao, H. Wang, J. Zheng, and H. Guo, "A semantic k-anonymity privacy protection method for publishing sparse location data," in *2019 Seventh International Conference on Advanced Cloud and Big Data (CBD)*, pp. 216–222, Suzhou, China, 2019.

[26] I. Santos, E. Coutinho, and L. Moreira, "K-anonymity technique for privacy protection: a proof of concept study," in *2019 Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pp. 391–396, Português (Brasil), 2019.

[27] Y.-H. Zhou, G.-H. Li, Y.-G. Yang, and W.-M. Shi, "Location privacy protection nearest neighbor querying based on Geohash," *Computer Science*, vol. 46, no. 8, pp. 212–216, 2019.

[28] S. Guochao, C. Guanghui, and W. Shaoxin, "Location privacy protection approach based on interval region," *College of Computer Science and Engineering*, vol. 56, no. 8, pp. 66–73, 2020.

[29] Y. Zhong, T. Wang, C. Gan, and X. Luo, "The location privacy preserving scheme based on Hilbert curve for indoor LBS," in *Advances in Swarm Intelligence*, Y. Tan, Y. Shi, and B. Niu, Eds., pp. 387–399, Springer, 2019.

[30] X. Sun, Y. Sun, Z. Xia, and J. Zhang, "The one-round multiplayer discrete Voronoi game on grids and trees," *Theoretical Computer Science*, vol. 838, pp. 143–159, 2020.

[31] Y. H. Zhang, Y. J. Gong, Y. Gao, H. Wang, and J. Zhang, "Parameter-free Voronoi neighborhood for evolutionary multimodal optimization," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 335–349, 2020.

[32] H. Kaplan, W. Mulzer, L. Roditty, P. Seiferth, and M. Sharir, "Dynamic planar Voronoi diagrams for general distance functions and their algorithmic applications," *Discrete & Computational Geometry*, vol. 64, no. 3, pp. 838–904, 2020.

[33] X. Zhao, D. Pi, and J. Chen, "Novel trajectory privacy-preserving method based on prefix tree using differential privacy," *Knowledge-Based Systems*, vol. 198, 2020.

[34] Y.-H. Zhou, G.-H. Li, Y.-G. Yang, and W.-M. Shi, "Location privacy preserving nearest neighbor querying based on Geohash," *Computer Science*, vol. 46, no. 8, 2019.

[35] K. Cao, Q. Sun, H. Liu, Y. Liu, G. Meng, and J. Guo, "Social space keyword query based on semantic trajectory," *Neurocomputing*, vol. 428, 2021.

[36] K. Xing, Y. Luo, X. Ning, and X. Zheng, "Location privacy protection algorithm based on Geohash encoding," *Computer Engineering and Applications*, vol. 55, no. 1, pp. 102–108, 2019.

[37] C. Y. Chow, M. F. Mokbel, and W. G. Aref, "The new Casper: query processing for location services without compromising privacy," *ACM Transactions on Database Systems*, vol. 34, 2009.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6690824, 21 pages
https://doi.org/10.1155/2021/6690824

Research Article

Hybrid Strategy of Multiple Optimization Algorithms Applied to 3-D Terrain Node Coverage of Wireless Sensor Network

Li-Gang Zhang[1], **Fang Fan**[1], **Shu-Chuan Chu**[1], **Akhil Garg**[2] and **Jeng-Shyang Pan**[1]

[1]*College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, 266590 Shandong, China*
[2]*State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, HuBei, China*

Correspondence should be addressed to Jeng-Shyang Pan; jengshyangpan@gmail.com

Received 5 November 2020; Accepted 10 July 2021; Published 10 August 2021

Academic Editor: Laurie Cuthbert

The key to the problem of node coverage in wireless sensor networks (WSN) is to deploy a limited number of sensors to achieve maximum coverage. This paper studies the hybrid strategies of multiple evolutionary algorithms, and applies them to the problem of WSN node coverage. We first proposed the hybrid algorithm SFLA-WOA (SWOA) based on Shuffled Frog Leaping Algorithm (SFLA) and Whale Optimization Algorithm (WOA). The SWOA algorithm combines the advantages of SFLA and WOA; that is, it retains the unique evolution model of WOA and also has the excellent co-evolution capability of SFLA. Secondly, using the mutation, crossover and selection operations of the differential evolution (DE) algorithm to further optimize this hybrid algorithm, the SWOA-based SFLA-WOA-DE (SWOAD) algorithm is proposed. In addition, the performance of SWOA and SWOAD has been tested by 30 benchmark functions in the CEC 2017 test set. Experimental results show that the optimization effects of these two algorithms are very outstanding. Finally, the simulation results show that the optimization algorithm proposed in this paper has a good effect on improving the signal coverage of WSN under the actual three-dimensional terrain.

1. Introduction

The Internet of Things makes use of local area networks or the Internet and other means of communication to achieve the interconnection of people, machines and things so as to realize the intelligent management of items and real-time perception of the environment [1]. WSN is one of the core technologies of the Internet of Things. It is also an important product of the integration of the information industry (computing, communications and sensors) in the new era. It has received extensive attention from various countries and organizations, and has formulated relevant strategic policies. For example, the U.S. Science Foundation (NSF) has developed a WSN research program to support research on relevant fundamental theories. The EU's sixth framework plan also emphasizes the importance of WSN and regards it as one of the hot areas for vigorous development in the future. Compared with traditional networks, WSN is low-cost, easy to deploy, has better fault tolerance and can be placed in any environment. The organizer can quickly build a fully functional WSN under limited time and conditions. Once the WSN has been set up, the maintenance and management work are carried out within the network and does not require much workforce. Therefore, the application field of WSN is very broad, and it can be used in military, modern industry and agriculture, environmental protection and other fields [2–4].

The rapid development of artificial intelligence also brings a variety of problems, and traditional calculation methods cannot solve them well. Intelligent evolutionary algorithms came into being and developed rapidly. At present, various evolutionary algorithms have been proposed, such as Genetic Algorithm (GA) [5–7], DE [8–10], Particle Swarm Optimization (PSO) [11–13], Artificial Bee Colony

(ABC) [14, 15], Multi-Verse Optimizer (MVO) [16, 17], Ant Lion Optimizer (ALO) [18], Grey wolf optimizer (GWO) [19], Cuckoo Search (CS) [20–22], Moth Flame Optimizer (MFO) [23, 24], Sine Cosine Algorithm (SCA) [25, 26], QUasi-Affine TRansformation Evolutionary (QUATRE) [27, 28], pigeon inspired optimization (PIO) [29, 30], Shuffled Frog Leaping Algorithm (SFLA) [31–35], Whale Optimization Algorithm (WOA) [36–39]. Meta-heuristic algorithms have received more and more attention from researchers due to their outstanding performance in solving optimization problems. They have been widely used in problems in transportation, wireless sensor networks, industrial production, intelligence system and other fields [40, 41]. But according to the No Free Lunch (NFL) theorem, there is no meta-heuristic algorithm that can be widely applied to various problems [42, 43]. In other words, optimization algorithms that achieve good performance on some problems may perform poorly on other problems. Therefore, new algorithms need to be proposed to solve increasingly complex problems. For example, propose a new heuristic algorithm, or improve the existing algorithm [44, 45], or combine two or more different algorithms to solve more complex problems [46, 47].

WSN is a network system composed of sensor nodes with sensing capabilities deployed in the detection area, and communicate through wireless communication. This emerging technology has brought a new way to obtain information and control management. And because WSN itself is very different from traditional networks, it brings a lot of challenges to people. In WSN, signal coverage can be defined as the ratio of the perceptible area to the entire area. The question of how to maximize network coverage for a given number of sensors is an important one. Intelligent evolutionary algorithms are also increasingly used to improve the coverage of WSN signals. For example, an intelligent calculation algorithm for enhancing black holes is proposed and used to solve the node coverage problem of wireless sensor networks under three-dimensional terrain [48]. An artificial bee colony algorithm with dynamic search strategy is proposed to solve the deployment problem of three-dimensional surface sensors and improve the signal coverage [49]. A genetic algorithm-based network coverage and optimization control strategy is proposed to solve the coverage problem of sensor nodes in three-dimensional terrain [50]. Therefore, this article attempts to mix WOA and SFLA to improve the performance of the original algorithm, and to deal with the problem of node coverage in WSN under 3D actual terrain.

The rest of this article is organized as follows. Related work introduced the principles of WOA, SFLA and DE, as well as the problem of WSN node coverage in a 3D actual environment. In Section 3, the process of mixing WOA and SFLA and the steps of using DE to optimize the hybrid algorithm are introduced. In Section 4, the performance of the proposed algorithm is tested, and the performance of the algorithm on 30 test functions is shown and analyzed. Section 5 introduces the application of the algorithm in WSN node coverage under actual terrain. Finally, the conclusion is given in Section 6.

2. Related Work

This section briefly introduces the principles of WOA, SFLA and DE and the problem of node coverage in WSN under actual terrain.

2.1. WOA. Mirjalili et al. were inspired by the humpback whale's spiral bubble net predation strategy and proposed a new heuristic whale optimization algorithm. WOA includes three location update models: encircling mode, bubble-net attacking mode, and searching mode. The WOA flowchart is shown in Figure 1.

2.1.1. Encircling Mode. In order to cooperate in predation, the whales share the location information of their prey, and then the whales approach the whale closest to the prey in the group. Update the current whale $\vec{X}$ according to the whale with the best position, and the update equations are as follows:

$$\vec{D} = \left| \vec{C} \bullet \overrightarrow{X^*}(t) - \vec{X}(t) \right| \tag{1}$$

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A} \bullet \vec{D} \tag{2}$$

Where t represents the current iteration number, $\vec{D}$ has different expressions at different stages, $\overrightarrow{X^*}$ is the whale with the best position so far, $\vec{X}(t)$ is the current whale position of the t-th generation, $\vec{X}(t+1)$ is the current whale position of the $(t+1)$-th generation, Operator $(\bullet)$ means to multiply item by item, $|\bullet|$ means to take the absolute value, $\vec{A}$ and $\vec{C}$ are coefficient vectors, The update equations of vectors $\vec{A}$ and $\vec{C}$ are as follows:

$$\vec{A} = 2\vec{a} \bullet \vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2 \bullet \vec{r} \tag{4}$$

$$a = 2\left(1 - \frac{t}{T}\right) \tag{5}$$

$\vec{r}$ is a random vector distributed in [0,1], $\vec{a}$ linearly decreases from 2 to 0 during the iteration process, so $\vec{A} \in [-2, 2]$, $\vec{C} \in [0, 2]$, T is the maximum number of iterations.

(1.1.1) Bubble-Net Attacking Mode

According to the spiral Equation (6), the current whale moves in a spiral motion to approach the prey and update its position.

$$\vec{X}(t+1) = \overrightarrow{D'} \bullet e^{bl} \bullet \cos(2\pi l) + \overrightarrow{X^*}(t) \tag{6}$$

$$\overrightarrow{D'} = \left| \overrightarrow{X^*}(t) - \vec{X}(t) \right| \tag{7}$$

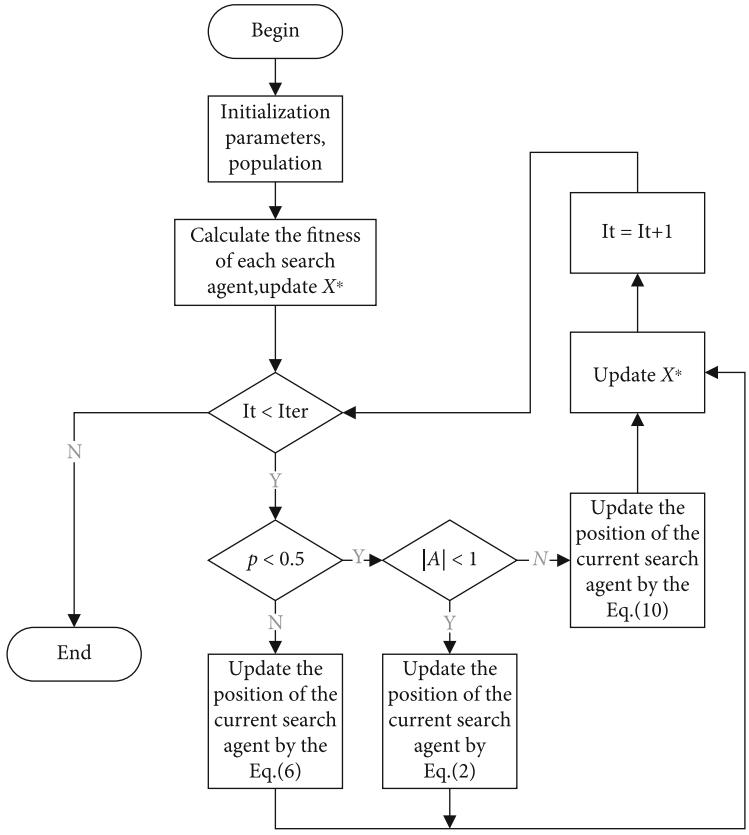


Figure 1: WOA flowchart.

$\overrightarrow{D'}$ denotes the distance between the current whale and the best positioned whale; b is a constant that defines the shape of a logarithmic spiral; and l is a uniformly distributed random number within [-1,1].

When whales spirally search for prey, they also shrink their encirclement. In order to simulate this behavior, the encircling prey and spiral search will be performed with the same probability. The location update equation is as follows:

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \vec{A} \bullet \vec{D}, & \text{if } p < 0.5 \\ \overrightarrow{D'} \bullet e^{bl} \bullet \cos(2\pi l) + \overrightarrow{X^*}(t), & \text{if } p \geq 0.5 \end{cases} \tag{8}$$

where p is a random number in the interval (0, 1).

2.1.2. Random Searching Mode. In order to improve the global search capability of whales, the current whale position is updated according to the randomly selected whales during the exploration phase. When $|A| < 1$, select the model that encircling; when $|A| \geq 1$, select the model of random search. The random search location update equations are as follows:

$$\vec{D} = \left| \vec{C} \bullet \overrightarrow{X_{\text{rand}}} - \vec{X} \right| \tag{9}$$

$$\vec{X}(t+1) = \overrightarrow{X_{\text{rand}}} - \vec{A} \bullet \vec{D} \tag{10}$$

Among them $\overrightarrow{X_{\text{rand}}}$ is a whale randomly selected from the current population.

2.2. SFLA. SFLA is a collaborative optimization algorithm proposed by Eusuff and Lansey et al. The idea of the hybrid leapfrog algorithm: When frogs hunt for food, they adjust their position through information exchange. First, the entire frog population is divided into multiple memeplexes, and each memeplex executes a local search strategy to adjust the position of the worst frog. When the memeplex iterates to the specified number of times, the memeplexes are combined and exchanged for information. The process of local search and the process of global information exchange are carried out cyclically until the end condition is met. The following are the steps of the hybrid leapfrog algorithm:

Step 1: Initialize the population and calculate the fitness value of each frog. Sort the population and record the individual P_b with the best position.

Step 2: The population containing F frogs is now divided into m memeplexes, so that there are n frogs in each memeplex, where n = F/m. If m =3, then the distribution principle is: the first frog is assigned to memeplex1, the second to memeplex2, the third to memeplex3, the fourth to memeplex1, ..., and so on.

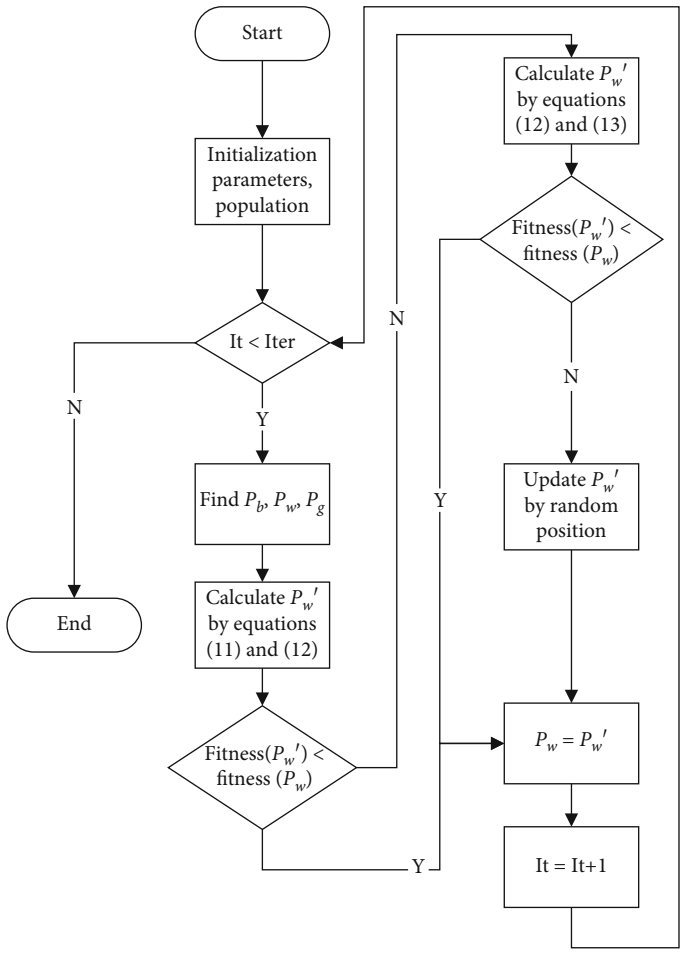


Figure 2: Local search flow chart of the hybrid leapfrog algorithm.

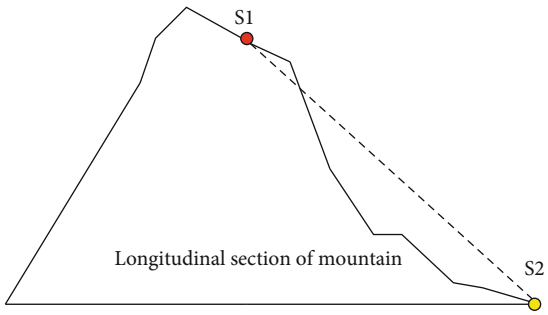


Figure 3: The communication process between sensor S1 and sensor S2.

Step 3: The local search process of the hybrid leapfrog algorithm is shown in Figure 2. Each memeplex evolves separately according to the following equations and Figure 2.

$$D_i = \text{rand} * (P_b - P_w) \tag{11}$$

$$P_w{}' = P_w + D_i, D_{\max} \geq D_i \geq -D_{\max} \tag{12}$$

$$D_i = \text{rand} * \left(P_g - P_w\right) \tag{13}$$

where P_b is the frog with the best position, P_w is the frog with

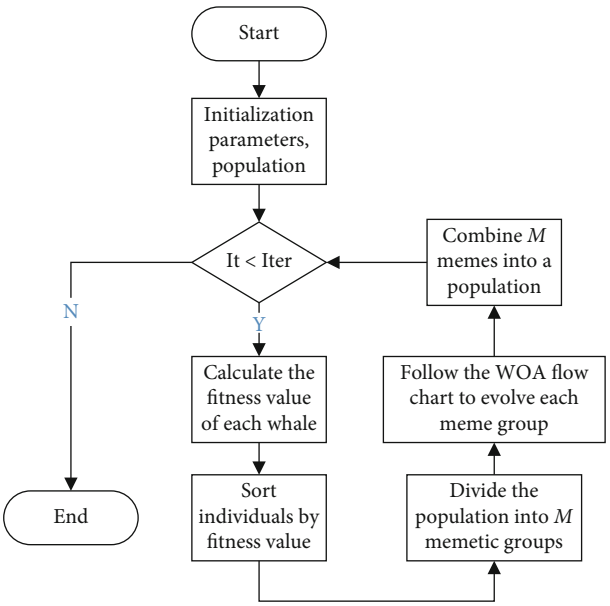


Figure 4: SWOA flow chart.

```
Initialize parameters; M is the number of memeplex; ne is the number of iterations for each memeplex; Iter is the maximum number of iterations.
for1 g =1: Iter do.
   Calculate the fitness value of each whale.
   Sort individuals by fitness value.
   Define memeplex, divide the population into M memeplexes.
   for2 each memeplex do.
      for3 n =1:Ne do.
           Update global optimal value.
        for4 each solution do.
          for5 each dimension do.
             Update Memeplex by equation (1) to equation (9).
          End for5.
        End for4.
      End for3.
      Combine M memes into a population.
   End for2.
End for1.
```

Algorithm 1: SWOA.

the worst position, rand() is a random number in [0-1], $P_w{}'$ is the adjusted position of the frog with the worst position, and P_g represents the best frog in the m memeplexes.

Step 4: After each memeplex evolves individually, it is reorganized into a population containing F frogs. Sort F frogs according to fitness value and update Pb.

Step 5: If the defined iteration conditions are met, the algorithm is terminated. Otherwise, go back to step 2.

2.3. De. The DE algorithm was proposed by Rainer Storn and Kenneth Price to solve the problem of Chebyshev polynomials. Differential evolution uses the three key operations of mutation, crossover and selection to continuously iterate to find the optimal value. First, the DE algorithm randomly selects several individuals in the population to perform mutation operations. Then crosses between the mutant individuals and the current individuals to obtain intermediates. Finally judges the pros and cons of the intermediates and the current individual, and selects individuals with good fitness values.

The mutation operation is based on all individuals in the population. Randomly select several individuals, one of which is the basis vector, and the other individuals make difference with each other to form a difference vector to construct a mutation operation. There are several combinations of basis vector and difference vector as follows:

$$m = x_{r1}^{t} + f\left(x_{r2}^{t} - x_{r3}^{t}\right) \tag{14}$$

where $r1$, $r2$, and $r3$ are unequal integers distributed in [1, Np], Np is the number of individuals in the population, t represents the current iteration number, m is the newly generated variant individual, f is the scale parameter for adjusting the solution size range, $f \in (0,1)$.

The crossover operation is to exchange the values of the mutated individual and the current individual in certain dimensions to form a new individual. The equation for binomial crossover is as follows:

```
for1 each solution do.
   Generate variant intermediates Y by Equation (14).
   The current solution is X.
   for2 each dimension of Y do.
      Generate new individuals U by Equation (15).
   End for2.
   if2 fitness(U) < fitness(X).
      X = U.
   End if2.
End for1.
```

Algorithm 2: DE pseudo code for step 6.

$$u_{i,j} = \begin{cases} y_{i,j}, & \text{rand}\ (0, 1) \le pCR \text{ or } d = \text{rand}\ ([1, numel(x)]) \\ x_{i,j}, & else \end{cases} \tag{15}$$

where pCR is the cross factor belonging to [0,1], rand $(\bullet)$ is the function of taking random values, d is the current dimension value, and $numel(\bullet)$ is the function of obtaining the total dimension of the individual.

The selection operation selects individuals who can enter the next generation population. If the fitness value of the new individual after mutation and crossover is better than the fitness value of the current individual, replace the current individual with the new individual, otherwise, keep the current individual.

2.4. 3D WSN Node Coverage. WSN has broad application prospects, and it has become one of the hot research fields today. The improvement of signal coverage in WSN has always been an important issue. The measurement of the WSN coverage can understand whether there is a blind spot for monitoring and communication, and then the

Table 1: Parameter settings for each related algorithm.

Name	Parameter
WOA	Np =40, Lb = -100, Ub =100, dim =30
SFLA	Np =40, Lb = -100, Ub =100, dim =30, Memeplex =5, M_it =25, Smax =10
SWOA	Np =40, Lb = -100, Ub =100, dim =30, Memeplex =5, M_it =25
SWOAD	Np =40, Lb = -100, Ub =100, dim =30, Memeplex =5, M_it =25, Beta_min =0.02, Beta_max =0.08, pCR =0.01

Table 2: Simulation Results of CEC 2017 Benchmark Function (The optimal value is marked by bold).

Functions	WOA		SFLA		SWOA		SWOAD	
Variable	Mean	Std	Mean	Std	Mean	Std	Mean	Std
F1	2.14E+10	5.14E+09	6.46E+10	8.46E+09	8.95E+09	2.81E+09	**2.85E+09**	**1.58E+09**
F2	3.12E+40	1.60E+41	4.97E+47	2.25E+48	1.72E+33	**4.16E+33**	**1.08E+33**	4.77E+33
F3	2.68E+05	7.76E+04	1.87E+05	4.61E+04	1.97E+05	**3.72E+04**	**1.52E+05**	4.59E+04
F4	4.84E+03	1.82E+03	3.12E+04	6.43E+03	1.42E+03	5.37E+02	**8.72E+02**	**3.60E+02**
F5	9.16E+02	5.80E+01	9.62E+02	4.06E+01	8.00E+02	**3.15E+01**	**7.65E+02**	3.45E+01
F6	6.88E+02	1.12E+01	6.97E+02	9.89E+00	6.70E+02	7.91E+00	**6.59E+02**	**6.88E+00**
F7	1.40E+03	8.81E+01	1.89E+03	1.83E+02	1.27E+03	**5.11E+01**	**1.22E+03**	7.31E+01
F8	1.11E+03	3.91E+01	1.15E+03	4.21E+01	1.03E+03	**2.08E+01**	**9.91E+02**	2.85E+01
F9	1.45E+04	5.55E+03	1.49E+04	2.65E+03	7.78E+03	**1.04E+03**	**6.63E+03**	1.07E+03
F10	8.34E+03	6.47E+02	8.07E+03	**5.23E+02**	6.75E+03	5.56E+02	**5.98E+03**	5.56E+02
F11	1.76E+04	8.70E+03	1.63E+04	4.40E+03	4.89E+03	1.38E+03	**3.23E+03**	**9.94E+02**
F12	1.82E+09	8.98E+08	1.91E+10	4.74E+09	4.31E+08	3.41E+08	**9.43E+07**	**6.22E+07**
F13	2.83E+08	2.26E+08	2.07E+10	6.73E+09	4.62E+06	**1.05E+07**	**4.54E+06**	2.30E+07
F14	3.52E+06	2.91E+06	1.13E+07	1.35E+07	**5.45E+05**	**4.64E+05**	7.12E+05	8.59E+05
F15	4.11E+07	3.90E+07	1.68E+09	1.10E+09	2.08E+06	2.18E+06	**3.65E+05**	**6.58E+05**
F16	4.68E+03	5.52E+02	5.41E+03	7.09E+02	3.65E+03	3.80E+02	**3.39E+03**	**3.46E+02**
F17	3.07E+03	3.16E+02	3.85E+03	1.43E+03	2.51E+03	2.21E+02	**2.37E+03**	**2.05E+02**
F18	3.96E+07	3.87E+07	5.77E+07	5.76E+07	2.96E+06	3.22E+06	**1.87E+06**	**2.24E+06**
F19	5.74E+07	4.26E+07	2.48E+09	1.96E+09	6.32E+06	8.39E+06	**9.67E+05**	**1.06E+06**
F20	3.02E+03	2.58E+02	3.18E+03	**1.21E+02**	2.69E+03	1.22E+02	**2.67E+03**	1.87E+02
F21	2.70E+03	5.99E+01	2.80E+03	5.35E+01	2.58E+03	**3.15E+01**	**2.55E+03**	4.12E+01
F22	9.08E+03	**1.39E+03**	9.73E+03	4.48E+02	**6.25E+03**	1.85E+03	6.31E+03	1.84E+03
F23	3.24E+03	1.18E+02	3.54E+03	1.08E+02	3.08E+03	**4.33E+01**	**3.05E+03**	8.38E+01
F24	3.35E+03	9.93E+01	3.86E+03	1.88E+02	3.18E+03	**6.77E+01**	**3.14E+03**	7.70E+01
F25	3.77E+03	2.42E+02	9.46E+03	1.48E+03	3.25E+03	1.17E+02	**3.09E+03**	**7.33E+01**
F26	9.34E+03	9.83E+02	1.04E+04	**8.53E+02**	7.67E+03	1.05E+03	**7.05E+03**	1.18E+03
F27	3.61E+03	1.82E+02	4.62E+03	3.50E+02	3.38E+03	**6.44E+01**	**3.35E+03**	7.44E+01
F28	5.02E+03	5.02E+02	9.78E+03	1.68E+03	3.92E+03	2.52E+02	**3.56E+03**	**1.46E+02**
F29	5.88E+03	7.14E+02	6.76E+03	1.24E+03	4.80E+03	3.66E+02	**4.60E+03**	**2.69E+02**
F30	1.70E+08	1.24E+08	1.77E+09	9.80E+08	2.36E+07	1.52E+07	**6.78E+06**	**4.89E+06**

deployment position of the sensor can be adjusted or the number of sensors can be increased to improve the coverage of the sensor signal. The high deployment density of sensor nodes will result in higher network coverage, but it will also cause redundancy of network coverage, resulting in a great waste of resources. In the case of a fixed number of sensors, the proper deployment location of WSN nodes will have a direct impact on network performance. Optimizing the location of wireless sensor nodes can reasonably allocate network resources and better complete environmental perception. At present, the sensor coverage strategy in the two-dimensional plane has achieved more results, applied to the three-dimensional space of the sensor coverage strategy is also gradually attracting the attention of scholars. For example, Adda Boualem et al. proposed a spider web strategy and applied it to area coverage in 3D wireless sensor networks using mobile sensor nodes [51]. Yu Xiang et al. proposed 3D space detection and coverage of wireless sensor network based on spatial correlation [52].

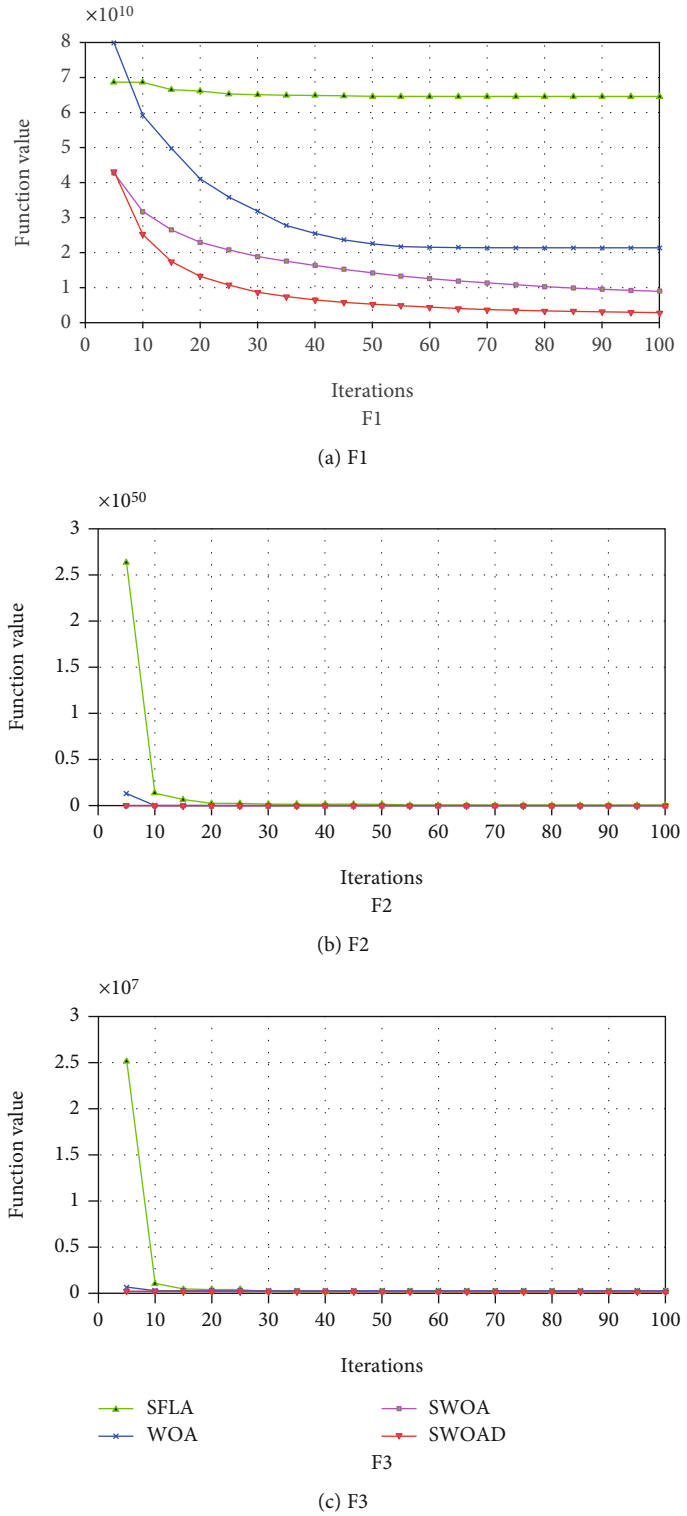
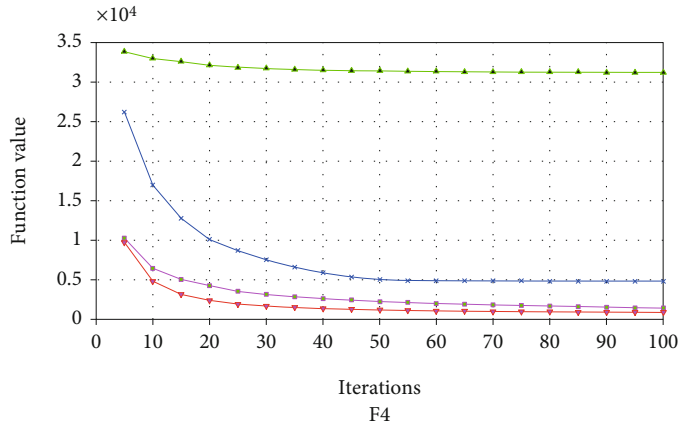
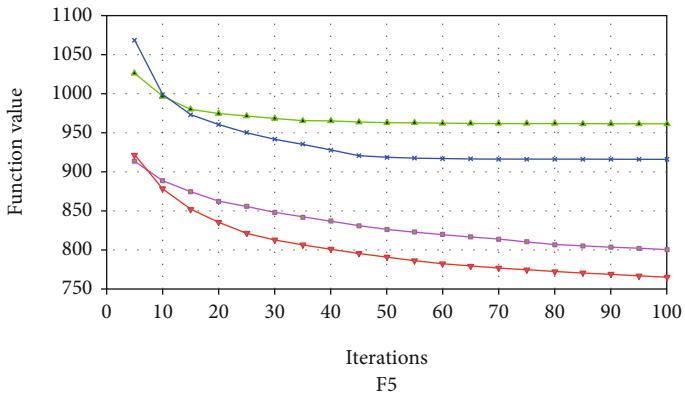


Figure 5: Convergence curves of unimodal functions.

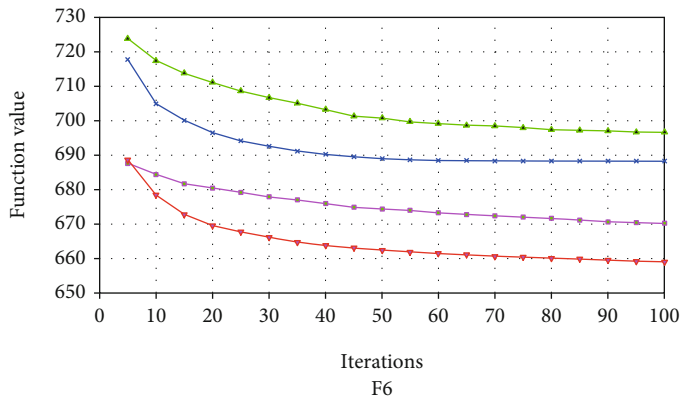
Deploying wireless sensors on an actual three-dimensional terrain rather than a two-dimensional plane requires additional considerations. Because the communication quality between the two sensors largely depends on the actual physical environment. Surrounding obstacles cause signal fading and obstruction. The communication process between sensor $S1(x1, y1, z1)$ and sensor $S2(x2, y2, z2)$ is shown in Figure 3. When the sensor S1 communicates with the sensor S2, the signal sent out is likely to be blocked by the protruding terrain between them, so that S1 and S2 will not be able to communicate, thereby reducing the coverage of the WSN signal. This paper uses Bresenham's line of sight



(a) F4

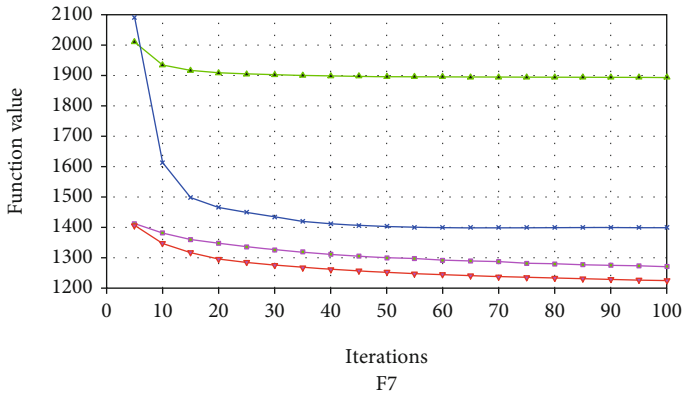


(b) F5

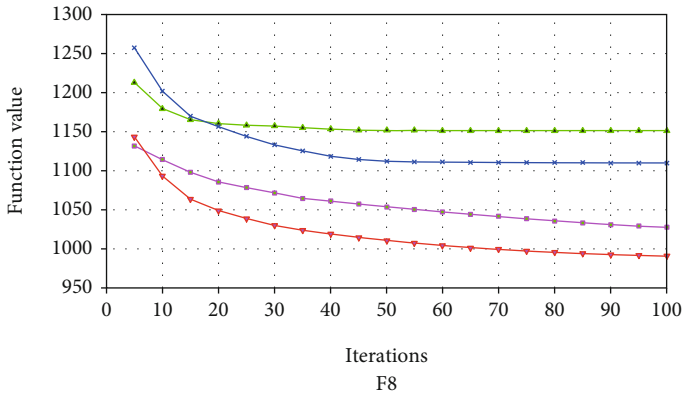


(c) F6

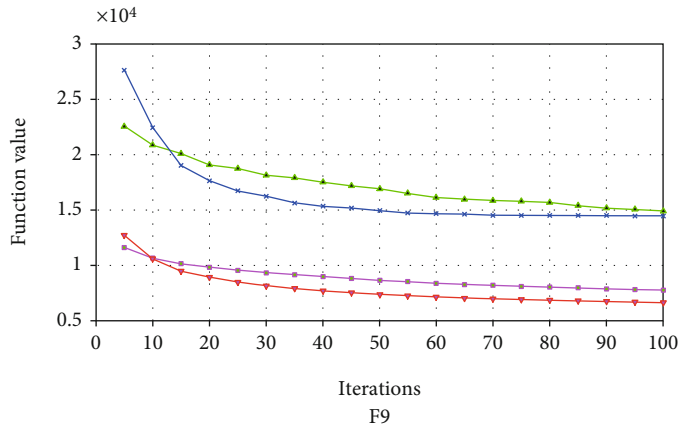
Figure 6: Continued.



(d) F7

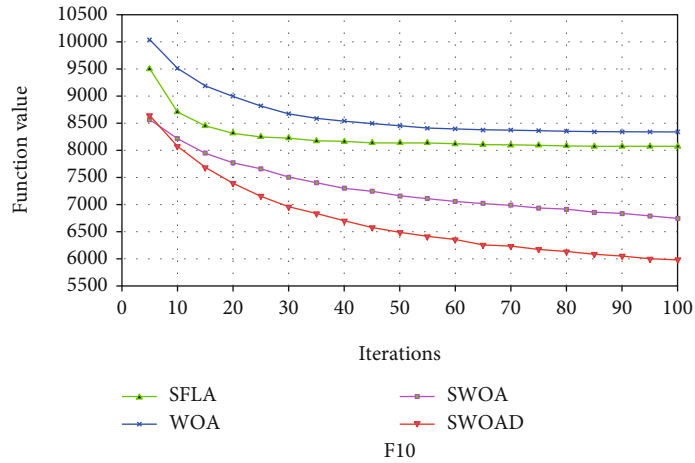


(e) F8



(f) F9

Figure 6: Continued.



(g) F10

Figure 6: Convergence curves of simple multimodal functions.

(LOS) algorithm to detect whether there is terrain obstruction between two sensors to affect their signal transmission [53]. In order to determine whether the communication between the target node S1 and the node S2 within the communication range is blocked, we choose the points between S1 and S2 on the actual terrain to make judgments. If the height of any one of the points is higher than the height of the connection between S1 and S2, the communication between S1 and S2 will be blocked, that is, the S1 node cannot communicate with the S2 node.

So to judge whether two nodes can communicate, first calculate the Euclidean distance D_s between sensor S1 and sensor S2 according to Equation (16).

$$D_s = \sqrt{(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2} \quad (16)$$

Then use the LOS algorithm to calculate whether there is an obstruction between the two sensors. Finally, a judgment is made according to Equation (17). If D_s is less than the coverage radius R_s of the sensor, and there is no obstruction between the two sensors, it means that they can communicate, and Communication (s1, s2) is equal to 1. Otherwise, if one of the conditions is not met, it is considered that sensor S1 and sensor S2 cannot communicate, and Communication (s1, s2) is equal to 0.

$$Communication(s1, s2) = \begin{cases} 1, & (D_s < R_s) \text{ and } No\ obstacle \\ 0, & else \end{cases} \quad (17)$$

D_s is the Euclidean distance between two sensors, and R_s is the coverage radius of the sensor signal.

3. Hybrid Strategy of SWOA and SWOAD

This section introduces the mixing process of WOA and SFLA, and the steps to optimize the mixing algorithm with DE.

3.1. Hybrid Strategy of SWOA. The WOA algorithm is a new type of bionic optimization algorithm with strong global optimization performance. However, in the later stage of the algorithm iteration, it still has the disadvantage of being easy to fall into the local optimal value, so the solution accuracy is low. The SFLA algorithm can realize global information exchange through the combination and sorting of memeplexes, so that the algorithm avoids falling into local convergence, but the early search speed is slow. The above shortcomings of the algorithm can be effectively solved by combining the WOA algorithm and the SFLA algorithm. The flow chart of SWOA algorithm is shown as in Figure 4.

SWOA's pseudo code is in Algorithm 1.

3.2. Hybrid Strategy of SWOAD. The SWOAD algorithm integrates the differential evolution algorithm into the SWOA algorithm, so that the whale population has a mechanism of mutation, crossover, and selection, which in turn enables SWOA to have stronger search capabilities. After each iteration of SWOA, the SWOAD algorithm first randomly selects a whale individual as the basis vector, and then combines with the difference between two randomly selected whale individuals to complete the mutation. Whether in the early or late stage of algorithm iteration, this strategy can enhance the ability of whales to jump out of the local optimum. Then the mutant intermediates are crossed with the target whale individuals, thereby increasing the diversity of the whale population; finally, a selection is made. If the fitness value of the newly generated individual is better than the target individual, the newly generated individual will replace the target individual. Otherwise, keep the target individual. SWOAD execution steps are as follows:

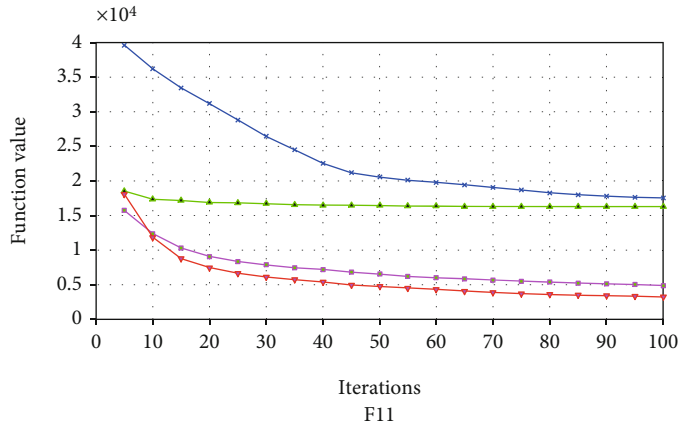
Step 1. Calculate the fitness of each individual.

Step 2. Sort the populations by fitness.

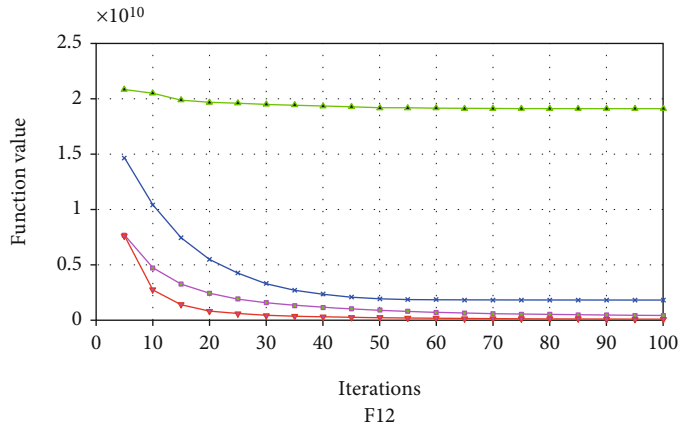
Step 3. According to the division rule, the population is decomposed into multiple memeplexes.

Step 4. Each memeplex evolves individually.

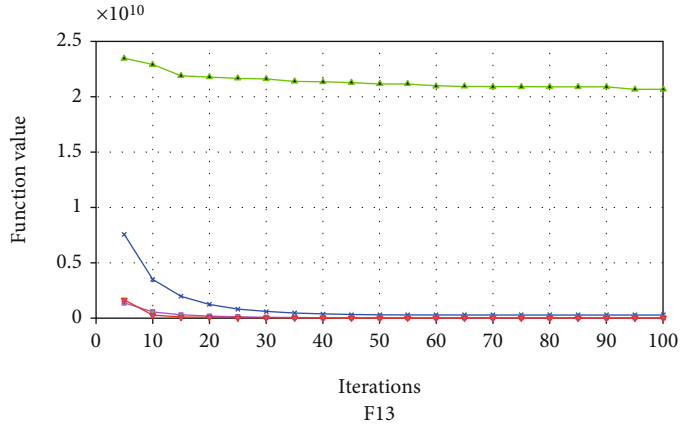
Step 4.1. Calculate the fitness value of each individual and Update global optimal value.



(a) F11

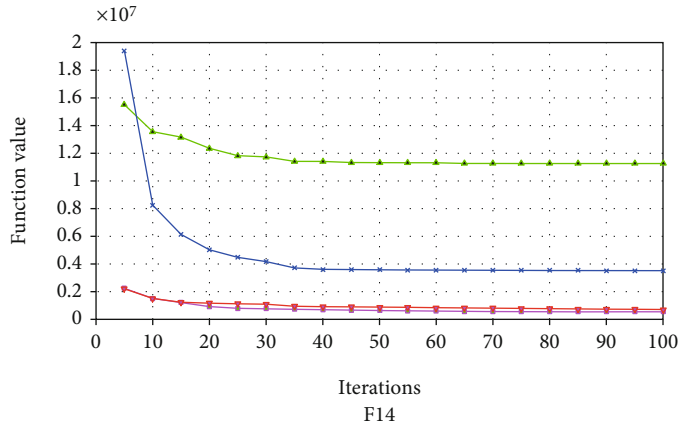


(b) F12

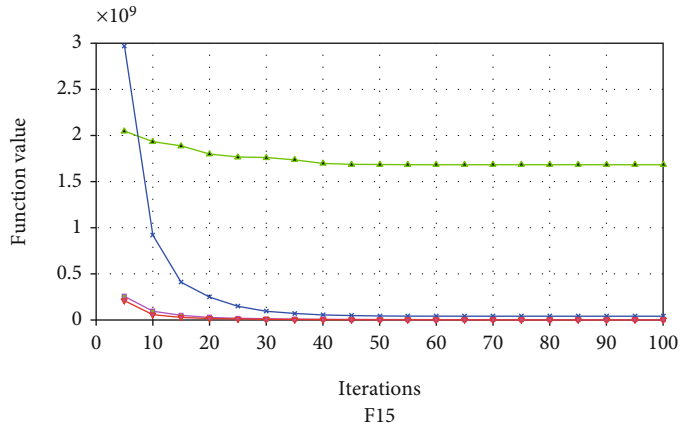


(c) F13

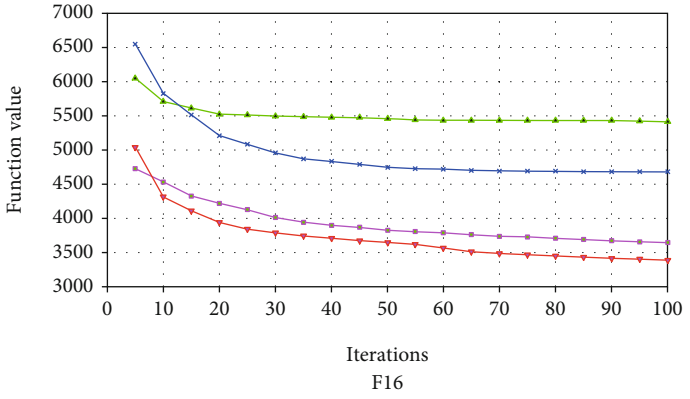
Figure 7: Continued.



(d) F14

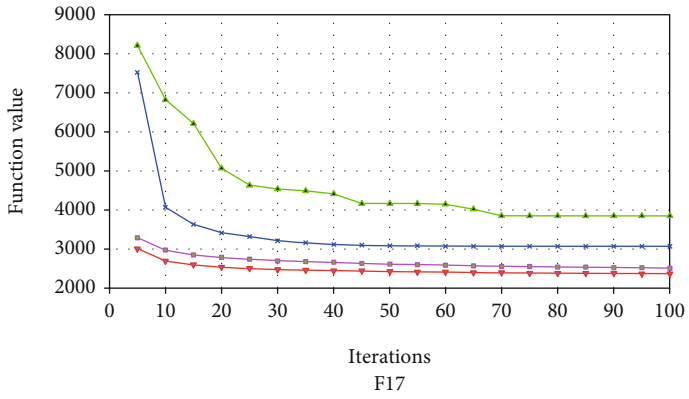


(e) F15

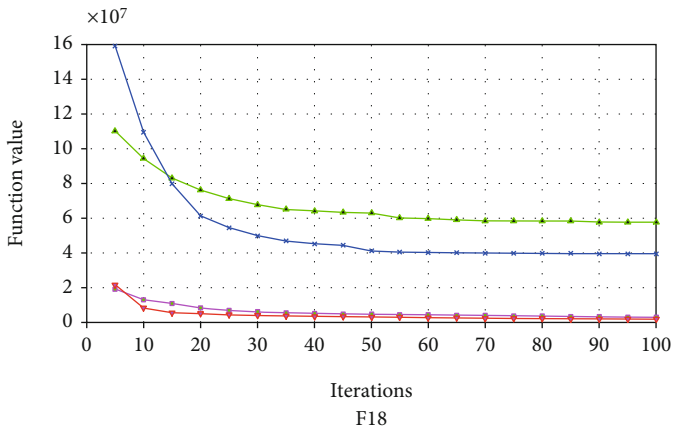


(f) F16

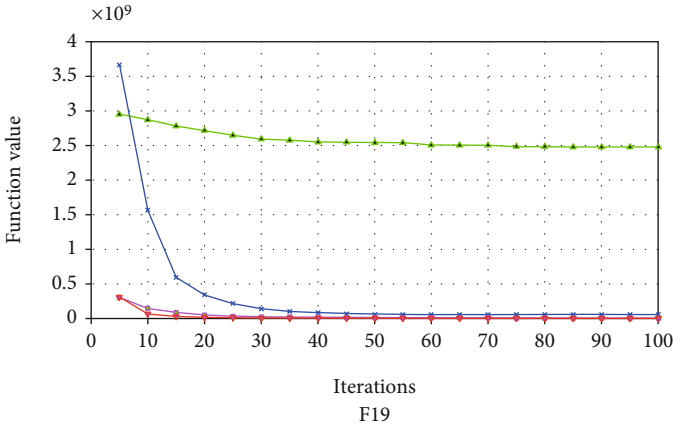
Figure 7: Continued.



(g) F17

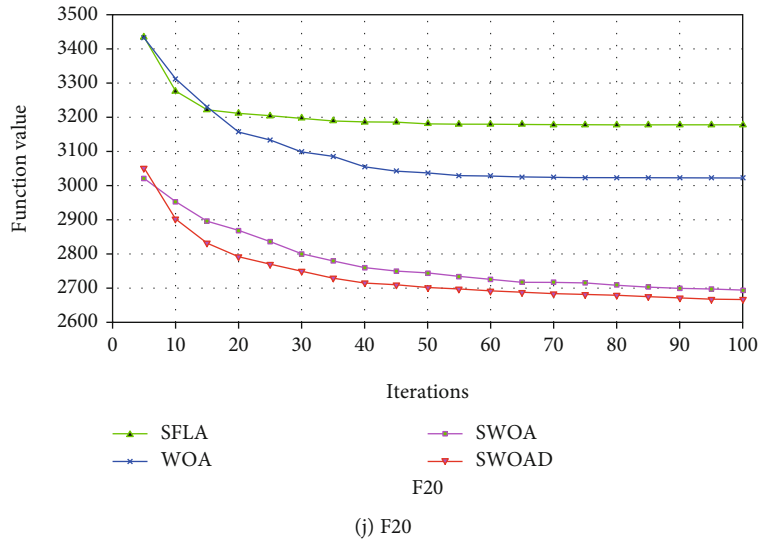


(h) F18



(i) F19

Figure 7: Continued.



(j) F20

Figure 7: Convergence curves of hybrid functions.

Step 4.2. Update the position of each frog by Equation (1) to Equation (10).

Step 4.3. Repeat 4.1–4.2 until the end conditions of the local search are met.

Step 5. Combine memeplexes.

Step 6. Performing variation, crossover and selection operations.

Step 7. Repeat step 1 to 6 until the end conditions are met.

The pseudo code of step 6 is in Algorithm 2.

4. Results and Analysis of the Experiment

In this section, we use 30 benchmark functions in CEC2017 to test the effectiveness of the proposed hybrid algorithm.

4.1. Parameter Configuration. To verify the results, we compared the hybrid algorithm with the original WOA and SFLA algorithms. Each algorithm performs 100 iterations on each benchmark function, and runs 30 times to average. The test parameters of the algorithm are given in Table 1. Table 2 shows the statistical results of the algorithm, including the mean (Mean) and standard deviation (Std).

From the data in the table, it can be seen that SWOA and SWOAD perform better than the original WOA and SFLA on the 30 test functions. SWOA performs better on functions F14 and F22, and SWOAD performs better on other functions. WOA has stability in function F22, SFLA has better stability in functions F10 and F20, SWOAD has excellent stability in function F1, F4, F6, F11, F12, F15, F16, F17, F18, F19, F25, F28, F29 and F30, and SWOA is more stable in other functions.

In order to further evaluate the performance of the algorithm, we use the convergence curve of the algorithm to evaluate the convergence speed and convergence ability of the optimized algorithm in this paper. The iterative curves of the algorithm on 30 test functions are shown in Figure 5–8. All algorithms have the same number of iterations on each function. The horizontal axis is the number of iterations of the function, and the vertical axis is the average of the fitness values of each function running 30 times.

4.2. Unimodal Functions. In Figure 5, SWOAD's convergence ability on function F1 is better than other algorithms. Each algorithm can find the optimal value on the function F2 and F3, but the improved two algorithms converge faster than the original algorithm.

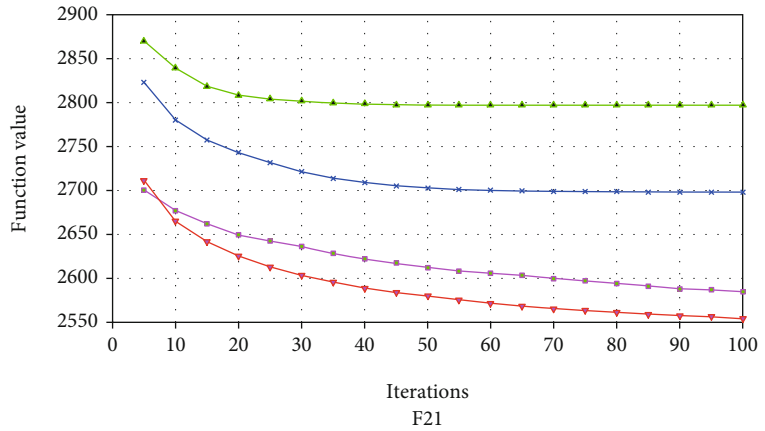
4.3. Simple Multimodal Functions. In Figure 6, the SWOAD algorithm has a faster convergence speed and stronger optimization ability than the WOA, SFLA and SWOA algorithms. The performance of SFLA in function F10 is better than that of WOA.

4.4. Hybrid Functions. In Figure 7, the convergence curves of SWOA and SWOAD on the function F13, F14, F15, F18 and F19 have very little difference. But the convergence performance of SWOAD on functions F11, F16, F17 and F20 are significantly better than SWOA. The performance of SFLA in function F11 is better than that of WOA.

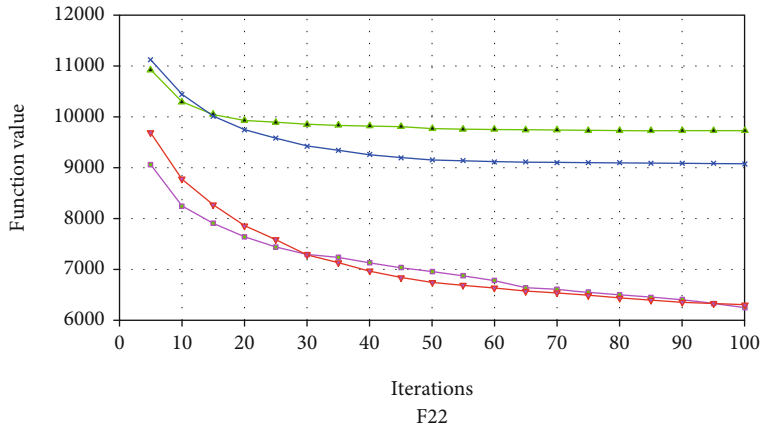
4.5. Composition Functions. In Figure 8, the convergence results of SWOA and SWOAD are almost the same in F22 and F30. Among other functions, SWOAD has the best performance, followed by SWOA, which is better than WOA and SFLA.

5. Application of Hybrid Algorithm in WSN Node Coverage under Real Terrain

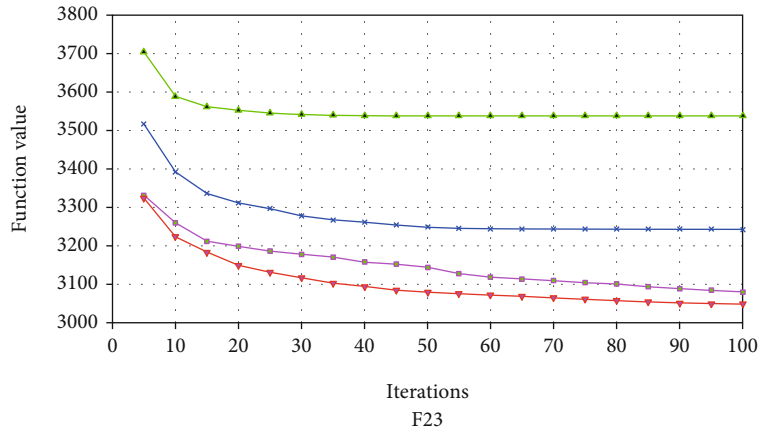
The 3-D actual terrain used in this simulation is the Dagong Island in Qingdao. Obtain topographic data of Dagong Island through satellite maps, and collect information about a coordinate point every ten meters in a prescribed area. The sensor nodes are deployed on this 3-D terrain. The terrain of Dagong Island is shown in Figure 9. When the simulation is performed on the actual terrain of Dagong Island, the initial sensor nodes are randomly generated and



(a) F21

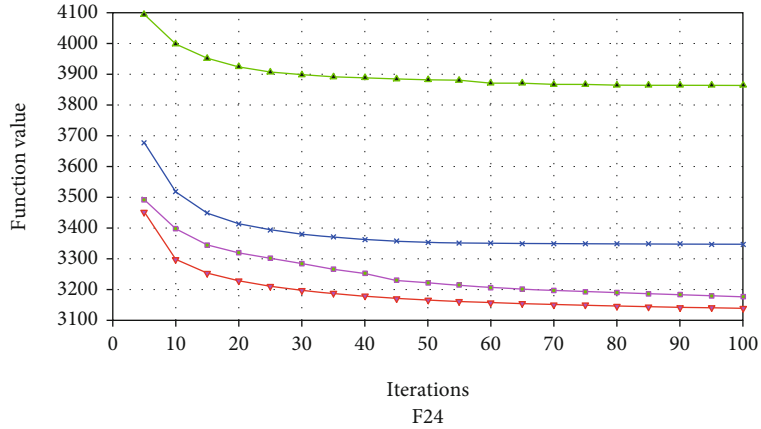


(b) F22

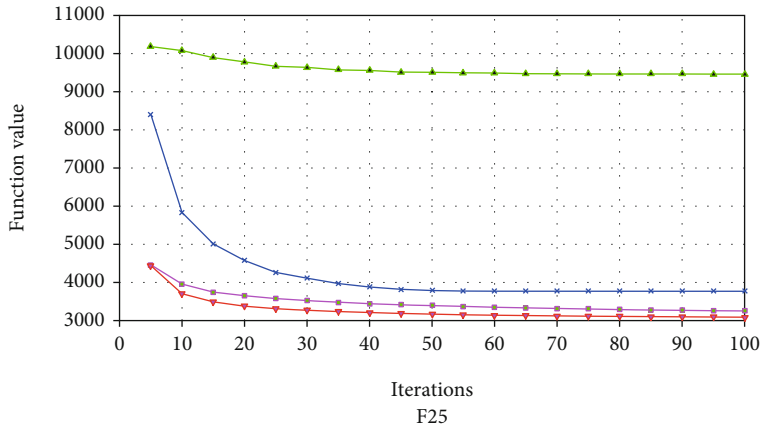


(c) F23

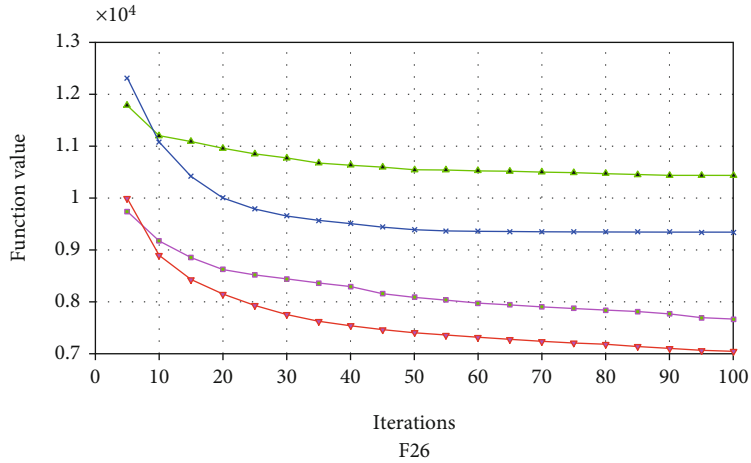
Figure 8: Continued.



(d) F24

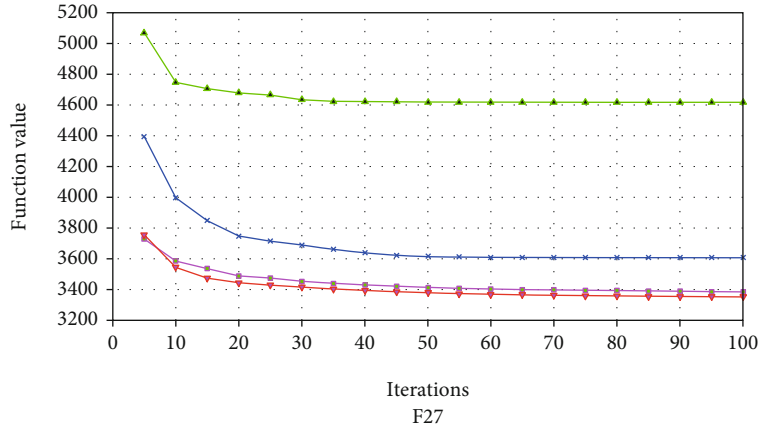


(e) F25

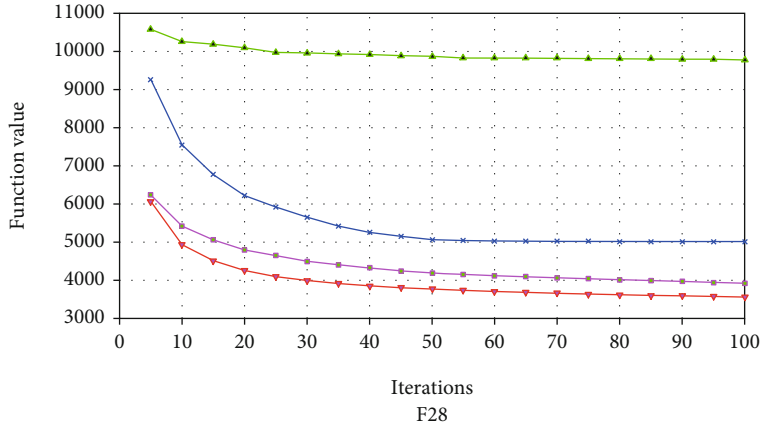


(f) F26

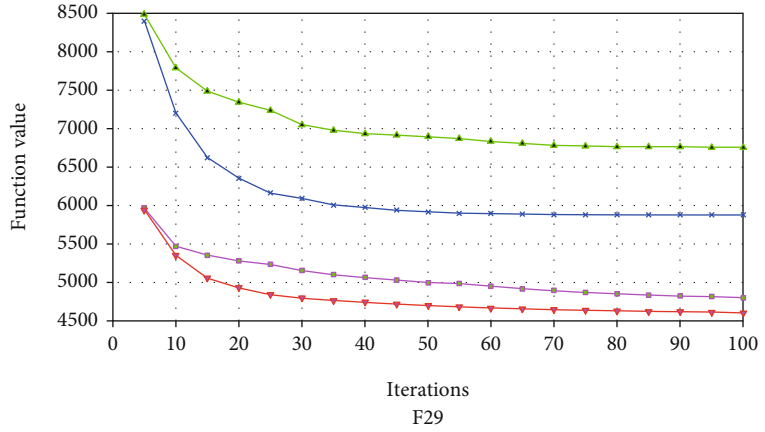
Figure 8: Continued.



(g) F27

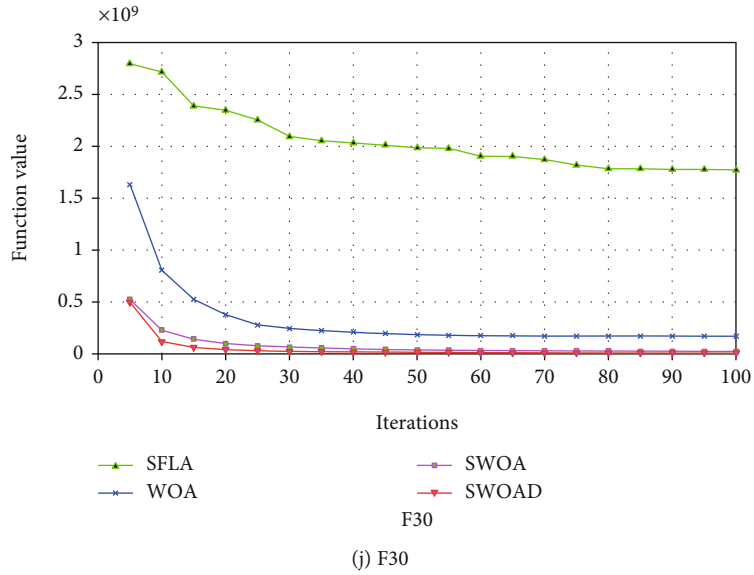


(h) F28



(i) F29

Figure 8: Continued.



(j) F30

Figure 8: Convergence curves of composition functions.

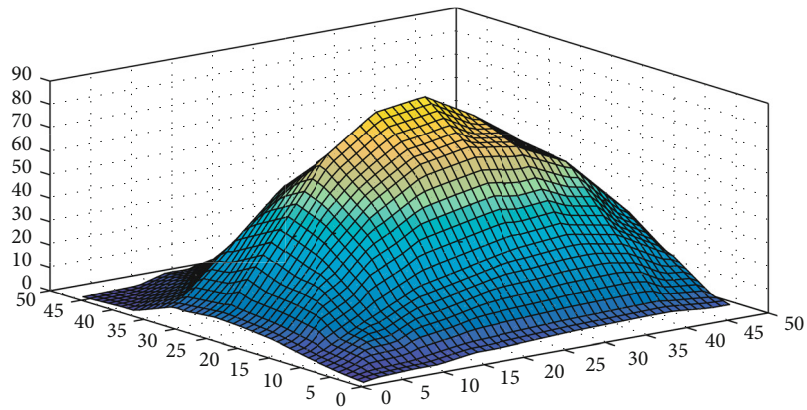


Figure 9: Topographic map of Dagong Island (unit is 10 meters).

optimized by hybrid algorithms to find a better position. To randomly generate the position of a sensor node, only the horizontal and vertical coordinates of the sensor node need to be randomly generated, and use the following method to determine the height through the available terrain data: If the randomly generated or optimized sensor node is at the intersection of the horizontal and vertical grid lines, the height value in the corresponding terrain data is the height of the sensor; Otherwise, the height of the grid intersection closest to this position is the height of the sensor node.

Set a matrix CMat, and determine the coordinate points that each sensor can cover by Equation (17), and set the coordinate points that the sensor can cover in CMat to 1. Calculate the coverage rate according to Equation (18).

$$Rate = sum(CMat(:) == 1)/(Xl * Yl) \quad (18)$$

where $sum(\bullet)$ is a sum function, Xl and Yl are horizontal and vertical coordinate lengths, respectively.

Table 3: Simulation Results of WSN node coverage (The optimal value is marked by bold).

Node number	Algorithm			
	WOA	SFLA	SWOA	SWOAD
30	0.5132	0.5583	0.5760	**0.5895**
40	0.6026	0.6621	0.6765	**0.6884**
50	0.6697	0.7428	0.7519	**0.7658**
60	0.7426	0.8012	0.8130	**0.8256**

Use the hybrid algorithm in this paper to optimize the position of the sensor, and improve the signal coverage as much as possible on the premise of a fixed number of sensors. The communication radius of the sensor is set to 5 m. Test the algorithm with 30, 40, 50, and 60 nodes, respectively. Run 30 times for each group of nodes and take the average. The results of the experiment are shown in Table 3. As the number of sensor nodes increases, the signal coverage also increases, and the coverage rate is the largest when 60 sensor

nodes are deployed, reaching 82.56%. It can also be seen that the hybrid algorithm is far superior to the performance of WOA and SFLA.

6. Conclusion

In this paper, WOA and SFLA are combined to form a new hybrid algorithm. The two algorithms cooperate with each other to form an organic whole. Compared with the performance of the two algorithms alone, the performance of the hybrid algorithm is better. Through mutual fusion, the algorithm can avoid falling into the local optimum in the process of finding the global optimum. And use DE to optimize the hybrid algorithm, which further improves the algorithm's convergence speed and optimization ability. In the experiments tested using the CEC 2017 benchmark function, the hybrid algorithm outperformed both WOA and SFLA. Finally, the hybrid algorithm is applied to the node coverage problem of wireless sensor network based on actual terrain. The simulation results show that the improved algorithm has achieved good results and increased the signal coverage of the wireless sensor network. There are many metaheuristic algorithms. In this article, we only use two algorithms for mixing. In the future, we may adopt some other algorithms [54–56] to get a hybrid approaches with better performance on WSN coverage problem.

Data Availability

(1) CEC 2017. (2) The actual terrain data is obtained from the LocaSpace Viewer software

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

References

[1] Y. Huang, X. Xue, and C. Jiang, "Semantic integration of sensor knowledge on artificial internet of things," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8815001, 8 pages, 2020.

[2] V. Bapat, P. Kale, V. Shinde, N. Deshpande, and A. Shaligram, "WSN application for crop protection to divert animal intrusions in the agricultural land," *Computers and Electronics in Agriculture*, vol. 133, pp. 88–96, 2017.

[3] W.-H. Nam, T. Kim, E.-M. Hong, J.-Y. Choi, and J.-T. Kim, "A Wireless Sensor Network (WSN) application for irrigation facilities management based on Information and Communication Technologies (ICTs)," *Computers and Electronics in Agriculture*, vol. 143, pp. 185–192, 2017.

[4] L. Q. V. Tran, A. Didioui, C. Bernier, G. Vaumourin, F. Broekaert, and A. Fritsch, "Co-simulating complex energy harvesting wsn applications: an in-tunnel wind powered monitoring example," *International Journal of Sensor Networks*, vol. 23, no. 2, pp. 100–112, 2017.

[5] D. Whitley, "A genetic algorithm tutorial," *Statistics and Computing*, vol. 4, pp. 65–85, 1994.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[7] H.-C. Huang, J.-S. Pan, Z.-M. Lu, S.-H. Sun, and H. M. Hang, "Vector quantization based on genetic simulated annealing," *Signal Processing*, vol. 81, no. 7, pp. 1513–1523, 2001.

[8] R. Storn and K. Price, "Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[9] J. Vesterstroem and R. Thomsen, "A comparative study of differential evolution particle swarm optimization and evolutionary algorithms on numerical benchmark problems," in *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*, vol. 2, pp. 1980–1987, Portland, OR, USA, 2004.

[10] J.-S. Pan, N. Liu, and S.-C. Chu, "A hybrid differential evolution algorithm and its application in unmanned combat aerial vehicle path planning," *IEEE Access*, vol. 8, pp. 17691–17712, 2020.

[11] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, Perth, WA, Australia, 1995.

[12] H. Wang, M. G Liang, C. L Sun, G. Zhang, and L. Xie, "Multiple-strategy learning particle swarm optimization for large-scale optimization problems," *Complex Intelligent Systems*, vol. 7, pp. 1–16, 2021.

[13] S. Qin, C. Sun, G. Zhang, X. He, and Y. Tan, "A modified particle swarm optimization based on decomposition with different ideal points for many-objective optimization problems," *Complex Intelligent Systems*, vol. 6, no. 2, pp. 263–274, 2020.

[14] D. Karaboga and C. Ozturk, "A novel clustering approach: artificial bee colony (ABC) algorithm," *Applied Soft Computing*, vol. 11, no. 1, pp. 652–657, 2011.

[15] D. Karaboga, *An Idea Based on Honey Bee Swarm for Numerical Optimization*, Tech. Rep. TR06, Department of Computer Engineering, Erciyes University, Turkey, Kayseri, 2005.

[16] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, "Multi-Verse optimizer: a nature-inspired algorithm for global optimization," *Neural Computing & Applications*, vol. 27, no. 2, pp. 495–513, 2016.

[17] X. Wang, J.-S. Pan, and S.-C. Chu, "A parallel multi-verse optimizer for application in multilevel image segmentation," *IEEE Access*, vol. 8, pp. 32018–32030, 2020.

[18] S. Mirjalili, "The ant lion optimizer," *Advances in Engineering Software*, vol. 83, pp. 80–98, 2015.

[19] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[20] X.-S. Yang and S. Deb, "Cuckoo search: recent advances and applications," *Neural Computing and Applications*, vol. 24, no. 1, pp. 169–174, 2013.

[21] P. C. Song, J.-S. Pan, and S.-C. Chu, "A parallel compact cuckoo search algorithm for three-dimensional path planning," *Applied Soft Computing*, vol. 94, p. 106443, 2020.

[22] 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.

[23] S. Mirjalili, "Moth-flame optimization algorithm: a novel nature-inspired heuristic paradigm," *Knowledge-Based Systems*, vol. 89, pp. 228–249, 2015.

[24] T.-T. Nguyen, H.-J. Wang, T.-K. Dao, J. S. Pan, T. G. Ngo, and J. Yu, "A scheme of color image multithreshold segmentation based on improved moth-flame algorithm," *IEEE Access*, vol. 8, pp. 174142–174159, 2020.

[25] S. Mirjalili, "SCA: a sine cosine algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 96, pp. 809–818, 2016.

[26] Q. Yang, S.-C. Chu, J.-S. Pan, and C.-M. Chen, "Sine cosine algorithm with multigroup and multistrategy for solving CVRP," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8184254, 10 pages, 2020.

[27] Z. Y. Meng and J. S. Pan, "QUasi-Affine TRansformation Evolution (QUATRE) Algorithm: a parameter-reduced differential evolution algorithm for optimization problems," in *Proceedings of the 2016 IEEE congress on evolutionary computation (CEC)*, pp. 4082–4089, Vancouver, Canada, July 2016.

[28] J. S. Pan, Z. Y. Meng, H. R. Xu, and X. Li, "QUasi-Affine-TRansformation Evolution (QUATRE) Algorithm: a new simple and accurate structure for global optimization," in *Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 657–667, Morioka, Japan, August 2016.

[29] H. Duan and P. Qiao, "Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning," *International Journal of Intelligent Computing and Cybernetics*, vol. 7, no. 1, pp. 24–37, 2014.

[30] 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.

[31] M. M. Eusuff and K. E. Lansey, "Optimization of water distribution network design using the shuffled frog leaping algorithm," *Journal of Water Resource Planning and Management*, vol. 129, no. 3, pp. 210–225, 2003.

[32] C. Liu, P. Niu, G. Li, Y. Ma, W. Zhang, and K. Chen, "Enhanced shuffled frog-leaping algorithm for solving numerical function optimization problems," *Journal of Intelligent Manufacturing*, vol. 29, no. 5, pp. 1133–1153, 2018.

[33] X. Li, L. Liu, N. Wang, and J. S. Pan, "A new robust watermarking scheme based on shuffled frog leaping algorithm," *Intelligent Automation & Soft Computing*, vol. 15, pp. l–15, 2011.

[34] A. Ouyang, X. Peng, Y. Liu, L. Fan, and K. Li, "An efficient hybrid algorithm based on hs and sfla," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 5, article 1659012, 2016.

[35] T. Niknam, M. R. Narimani, and R. Azizipanah-Abarghooee, "A new hybrid algorithm for optimal power flow considering prohibited zones and valve point effect," *Energy Conversion and Management*, vol. 58, pp. 197–206, 2012.

[36] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, no. 5, pp. 51–67, 2016.

[37] J.-S. Pan, J.-L. Liu, and E.-J. Liu, "Rank-based whale optimization algorithm for solving parameter optimization of solar cells," *International Journal of Modeling and Optimization*, vol. 9, no. 4, pp. 209–215, 2019.

[38] I. N. Trivedi, P. Jangir, A. Kumar, N. Jangir, and R. Totlani, "A novel hybrid PSO–WOA algorithm for global numerical functions optimization," in *Advances in Computer and Computational Sciences, vol. 554*, Springer, Singapore, 2017.

[39] A. Selim, S. Kamel, and F. Jurado, "Voltage Profile Improvement in Active Distribution Networks Using Hybrid WOA-SCA Optimization Algorithm," in *2018 Twentieth International Middle East Power Systems Conference (MEPCON)*, pp. 1064–1068, Cairo, Egypt, 2018.

[40] 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.

[41] X. Xue and J. S. Pan, "An overview on evolutionary algorithm based ontology matching," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, pp. 75–88, 2018.

[42] 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.

[43] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204, Athens, Greece, 2011.

[44] J. Zhuang, H. Luo, T.-S. Pan, and J.-S. Pan, "mproved flower pollination algorithm for the capacitated vehicle routing problem," *Journal of Network Intelligence*, vol. 5, no. 3, pp. 141–156, 2020.

[45] 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.

[46] J.-S. Pan, F. Fan, S.-C. Chu, Z. du, and H. Q. Zhao, "A Node Location Method in Wireless Sensor Networks Based on a Hybrid Optimization Algorithm," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8822651, 14 pages, 2020.

[47] P. Hu, J.-S. Pan, S.-C. Chu, Q.-W. Chai, T. Liu, and Z.-C. Li, "New hybrid algorithms for prediction of daily load of power network," *Applied Sciences*, vol. 9, no. 21, p. 4514, 2019.

[48] J.-S. Pan, Q. W. Chai, S.-C. Chu, and N. X. Wu, "3-D Terrain Node Coverage of Wireless Sensor Network Using Enhanced Black Hole Algorithm," *Sensors*, vol. 20, no. 8, p. 2411, 2020.

[49] H. Yang, X. Li, Z. Wang, W. Yu, and B. Huang, "A novel sensor deployment method based on image processing and wavelet transform to optimize the surface coverage in WSNs," *Chinese Journal of Electronics*, vol. 25, no. 3, pp. 495–502, 2016.

[50] F. Lin, Z. Sun, and T. Qiu, "Genetic algorithm-based 3d coverage research in wireless sensor networks," in *International Conference on Complex Intelligent and Software Intensive Systems (CISIS)*, pp. 623–628, Taichung, Taiwan, 2013.

[51] A. Boualem, Y. Dahmani, C. D. Runz, and M. Ayaida, "Spider-web strategy: application for area coverage with mobile sensor nodes in 3D wireless sensor network," *International Journal of Sensor Networks*, vol. 29, no. 2, pp. 121–133, 2019.

[52] Y. Xiang, Z. Xuan, M. Tang, J. Zhang, and M. Sun, "3D space detection and coverage of wireless sensor network based on spatial correlation," *Journal of Network and Computer Applications*, vol. 61, pp. 93–101, 2016.

[53] S. Temel, N. Unaldi, and O. Kaynak, "On deployment of wireless sensors on 3-D terrains to maximize sensing coverage by utilizing cat swarm optimization with wavelet transform," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 1, pp. 111–120, 2014.

[54] X. Xue, J. Chen, J. Liu, and D. Chen, "Matching biomedical ontologies through compact evolutionary simulated annealing

algorithm," in *In International Conference on Genetic and Evolutionary Computing*, pp. 661–668, Singapore, 2018.

[55] J.-S. Pan, Z. Meng, S.-C. Chu, and H.-R. Xu, "Monkey king evolution: an enhanced ebb-tide-fish algorithm for global optimization and its application in vehicle navigation under wireless sensor network environment," *Telecommunication Systems*, vol. 65, no. 3, pp. 351–364, 2017.

[56] Q.-W. Chai, S.-C. Chu, J.-S. Pan, and W.-M. Zheng, "Applying adaptive and self assessment fish migration optimization on localization of wireless sensor network on 3-d terrain," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 11, no. 2, pp. 90–102, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9977884, 12 pages
https://doi.org/10.1155/2021/9977884

Research Article

An Improved Algorithm Based on Fast Search and Find of Density Peak Clustering for High-Dimensional Data

Hui Du, Yiyang Ni, and Zhihe Wang

The School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China

Correspondence should be addressed to Hui Du; duhuiywy@nwnu.edu.cn

Received 1 April 2021; Revised 1 June 2021; Accepted 6 July 2021; Published 27 July 2021

Academic Editor: Xingsi Xue

The find of density peak clustering algorithm (FDP) has poor performance on high-dimensional data. This problem occurs because the clustering algorithm ignores the feature selection. All features are evaluated and calculated under the same weight, without distinguishing. This will lead to the final clustering effect which cannot achieve the expected. Aiming at this problem, we propose a new method to solve it. We calculate the importance value of all features of high-dimensional data and calculate the mean value by constructing random forest. The features whose importance value is less than 10% of the mean value are removed. At this time, we extract the important features to form a new dataset. At this time, improved t-SNE is used for dimension reduction, and better performance will be obtained. This method uses t-SNE that is improved by the idea of random forest to reduce the dimension of the original data and combines with improved FDP to compose the new clustering method. Through experiments, we find that the evaluation index NMI of the improved algorithm proposed in this paper is 23% higher than that of the original FDP algorithm, and 9.1% higher than that of other clustering algorithms (K-means, DBSCAN, and spectral clustering). It has good performance in high-dimensional datasets that are verified by experiments on UCI datasets and wireless sensor networks.

1. Introduction

In our daily life, when we are faced with a problem that we do not have an accurate standard to classify. Cluster analysis is an effective means to judge and analyze. It has applications in many fields, such as finance, medical, and image. The cluster algorithm divides elements into clusters according to calculated mathematical characteristics. Although many clustering algorithms have been studied and discussed, there is no agreement on the definition of clustering. Speaking of clustering algorithm, the first thing we have to mention is K-means [1]. K-means is an efficient and concise algorithm, which has been discussed by many researchers. It only needs to set the number of clusters to calculate clustering results to users. But these methods cannot detect clusters for nonspherical data [2]. DBSCAN [3] is also a classic and effective algorithm. It is a representative clustering algorithm based on density. DBSCAN is an algorithm that is in line with the spatial distribution of data and the consistency of data density. The author creatively defines two parameters to constrain and control the generation of clusters. But these two parameters also lead to the effect of the DBSCAN algorithm seriously affected by the parameters. It has good robustness [4] to outliers and can even detect outliers. Spectral clustering [5] is a method of clustering data points by means of discrete mathematics. The algorithm constructs an undirected graph and judges and analyzes the properties of the undirected graph. The most ingenious and important part of this algorithm is to construct Laplacian matrix. To construct Laplacian matrix, Laplacian matrix needs to calculate the similarity matrix. Firstly, the similarity matrix will use full connection mode. There are many kernel functions to measure the relationship between points. The average effect of Gaussian kernel function is the best among many kernel functions. Secondly, the Laplacian matrix is constructed by calculating adjacency matrix and degree matrix through similarity matrix. Finally, it needs to get the eigenvector of Laplacian. According to eigenvalues and eigenvectors, we can use other clustering algorithms to complete the clustering task. Balanced iterative reducing and clustering using

hierarchies (BIRCH) [6] is method by tree structure to cluster quickly. Birch algorithm is to form a clustering feature (CF) tree. By calculating the similarity of the dataset, the most similar sample data points in the dataset are combined, and the process is iterated [7]. Fuzzy clustering [8] is a new clustering algorithm with the development of the automation control field. In the theory of fuzzy mathematics, the concept of membership degree is mentioned for the first time. This theory greatly promotes its development. The main solution of the algorithm is to introduce membership degree to optimize or even solve problem when a point cannot be effectively identified. Affinity propagation (AP) is a clustering algorithm using a voting mechanism with good results. The idea of AP algorithm is very interesting, which is illustrated by examples in our life. For example, there are many commodities and many customers. If a commodity can be selected by customers, it must have its own strength. That is to say, it must have enough attraction. Secondly, it is the direct word-of-mouth of customers. One customer says yes, and further recommends the next customer to buy it, and so on. In this case, the commodity is like the cluster center to be selected, and the customer is the point attached to the cluster center. To sum up, the core of AP clustering algorithm is to calculate the attribution matrix and attraction matrix.

After introducing FDP, there are many papers that solve FDP's problem. This paper proposes a method [9] that uses the characteristics of data point density distribution to cluster efficiently and quickly. The algorithm ingeniously designs a set of calculation method and selects the point with the highest density in a certain area as the center. Noncentral point assignment is based on the class of the nearest point whose density is larger than itself. Outliers are found and excluded from the analysis. Regardless of the shape of the cluster and the dimension of the embedded space, clustering will be recognized.

However, FDP cannot solve many data features or high-dimensional data effectively, some nonmain features will interfere and affect the performance of the algorithm, and the scientists have made a series of improvements. Among them, dimension reduction is the first choice. A new fast hybrid dimension reduction method [10] is proposed. It innovatively proposes a method that combines multiple feature extraction methods. In this way, the interference of nonmain features can be reduced and the efficiency can be improved. The Fisher score and feature selection based on information gain are used to remove nonmain features. In this way, the interference of nonmain features is reduced, and the clustering effect is significantly improved. At present, there are many ways to reduce dimension, and principal component analysis (PCA) [11] is the best one. PCA creatively maps high-dimensional features to the low-dimensional features, so that the new low-dimensional data is an orthogonal matrix, which is also called the main component. Because it is a low-dimensional feature reconstructed on the basis of the original high-dimensional feature, the original feature is kept while the dimension is reduced. In conclusion, we compare with the PCA in experiments.

The purpose of studying the clustering algorithm is to serve people's life. In this paper [12], the popular concept of topology is used. In topology, manifold learning is the key to this paper. Manifold learning is currently a popular dimension reduction method. Let us briefly introduce manifold learning. For example, a piece of paper can be seen as two-dimensional when it is tiled. But when it is kneaded into a mass, it can be regarded as three-dimensional. For any point on the tissue, whether it is kneaded into a ball or flat, its relative position has not changed. If we can reduce the dimension smoothly and keep the feature unchanged. At the same time, the index matrix of the discrete clustering optimization process is easily affected by noise. To solve matters, a new clustering way was proposed in this paper. It combined local adaptive subspace learning and knowledge clustering to mine discriminant information adaptively.

Aiming at the practical problems of sample data nonlinearity and high dimension in complex system evaluation or prediction, in paper [13], they all used a memetic algorithm model [14] to achieve dimensionality reduction and achieved good results.

In another paper, it proposed a clustering algorithm for high-dimensional stream data. This algorithm introduced dimension reduction into the framework of stream clustering. When the new data arrives, for the sake of finding the local shadow space, there is necessary processing of the disordered new data. It is very necessary to reduce the high dimension to the low dimension. The algorithm innovatively used the unsupervised linear discriminant analysis (LDA) to process the data, so as to find the local shadow space. The obtained local subspace will maximally separate the adjacent microclusters from the incident point. Introduction points are admeasured to the microclusters in the projection space, which can be improved by this method.

Facing high-dimensional data, we proposed a method to combine FDP to achieve the performance of FDP on high-dimension data.

At this moment, let us discuss the practical application of clustering. When it comes to practical application, we have to mention the Internet of Things (IoT). It is often used in the intelligent world. As a sensing layer, wireless sensor networks are composed of many sensor networks. The sensor is the most important part of Internet of Things. Sensors are like human facial features to perceive the world. Sensors constantly provide information for users to facilitate their production and life. Therefore, to obtain a stable network, the energy control of the sensor must be optimized first, so as to extend the life of each sensor in the network. Sensor in the network is like a data point in space. At this time, each data point has a series of characteristics, such as relative distance and energy consumption. How to better control the energy consumption of wireless sensor is the difficult problem.

In this paper, [15] creatively introduced the semantic relationship between sensors to promote the effect of the whole network. Therefore, it proposed the new sensor ontology integration technology by introducing such a mechanism which is called the debate mechanism (DM). The purpose of this method is to extract sensor ontology alignment [16, 17].

In this way, the communication ability between wireless sensors is strengthened, and the performance of the whole network is improved. The global factor in the algorithm is calculated by the correctness factor. The local factor is calculated through the debate mechanism. By getting the global factor and the local factor, the judgment factor can be obtained by combining them. The judgment factor is used to achieve ontology alignment. By this means, the effect of the whole wireless sensor network can be improved. Inspired by this article, we finally apply the improved FDP to wireless sensor networks and do a comparative experiment to show the performance.

2. Related Work

2.1. Introduction to FDP. ρ_i and d_{ij} are important values of FDP. Equations (1) and (2) are designed to evaluate local density.

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \tag{1}$$

$$\chi(x) = \begin{cases} 1, x < 0, \\ 0, x = 0, \end{cases} \tag{2}$$

where ρ_i is used to represent the local density. The degree of similarity between point i and point j is expressed by d_{ij}. It is generally measured by Euclidean distance. For Equation (1), d_c is usually set to 1%-2% of the total number of samples (the total number of data points). Although the original text does not specifically point out, the setting of d_c needs to be set by the user, and the impact on the algorithm is very huge. Partly, the setting of d_c is also very difficult. This parameter has a great influence on the algorithm. Moreover, if the setting is not correct, the expected effect will not be obtained, and it is mistaken for the performance of the algorithm itself. Its value setting is also a very popular direction, and its value setting can effectively improve the performance of FDP.

Secondly, δ_i is the nearest Euclidean distance at all points with a greater density than itself. The decision graph is set by ρ_i and δ_i as the x-axis and y-axis of the coordinate axis, respectively. The larger and larger sample points are selected as the center of the class cluster in the decision graph. It needs to check manually the region according to the generated decision graph, and the point in the selected region is the cluster center point. The cluster of each noncenter sample point is the cluster of the nearest sample point higher than the point in the neighborhood.

2.2. Introduction to t-SNE. Among many dimensionality reduction algorithms for high-dimensional data, the stochastic neighbor embedding (SNE) [18] is a very special one. The idea of its algorithm is very clever and simple, but it contains a lot of probability theory. It is very similar for the popular learning mentioned above. It is about flattening a spatial dimension into a plane. The algorithm first describes the distribution of each point; how to measure the distribution is a key. There are many ways to measure; here, we choose Euclidean distance first. According to the research, different measurement methods have a great impact on the performance of the algorithm. We consider two data points x_i and x_j which are high-dimensional data. The conditional probability represents the degree to which data x_j is a neighbor of data x_i by $p_{j|i}$. We can define $p_{j|i}$ in this way.

$$P_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)}, \tag{3}$$

where σ_i is a statistical constant. First, we get the Gaussian distribution x_i which is centered and then calculate its variance.

The essence of dimension reduction is not to change the nature of data points and the relationship between data points. Low-dimensional data can keep as many features as possible. y_i and y_j are the data in low-dimensional corresponding to high-dimensional x_i and x_j separately. Through such a mapping relationship is to establish the dimensionality reduction process. The conditional probability density of a low-dimensional is defined just like that of a high-dimensional by $q_{j|i}$.

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right)}. \tag{4}$$

It is obvious that two conditional probability densities are obtained by calculation, and how to deal with and use these two values will be the key of this algorithm. If the high- and low-dimensional distributions are consistent, then the two conditional probabilities will be equal. Then, our goal is relatively clear. When the two conditional probabilities are equal or the difference is very small, the effect of dimension reduction will be perfect. The author of SNE introduced the Kullback-Leibler (KL) divergence distance to solve this problem.

$$\sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{5}$$

The dimension reduction effect of SNE is catastrophic for clustering. The clustering algorithm cannot cluster the reduced data effectively. Such dimension reduction is meaningless for clustering. Because SNE pays more attention to the local structure and ignores global structure. Having a certain impact, there is also congestion problem when using symmetric SNE.

The performance of t-distribution and Gaussian distribution is similar without interference data. But when there are outliers in the sample, there are inconsistencies. The simulation result of Gaussian distribution is not as good as that of t-distribution. t-distribution can keep the internal structure of the original data unchanged, and the variance is small. t-distribution can keep and show the characteristics of the original data better. At this time, the author can

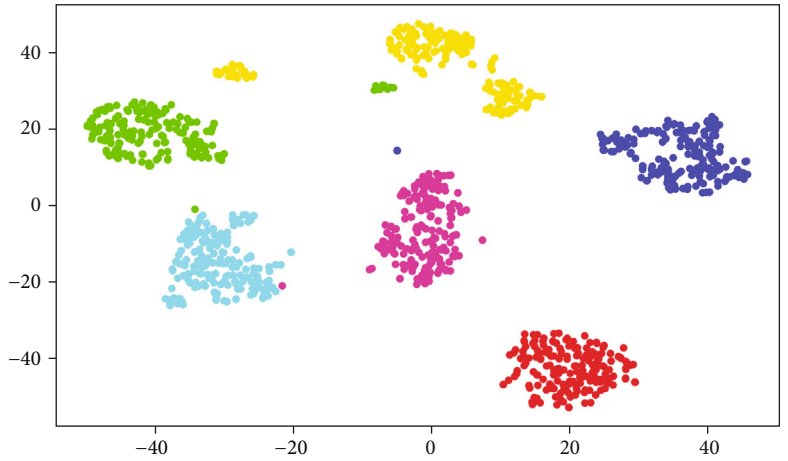


Figure 1: t-SNE for 0-5 digits. We select digits dataset 0-5 and use t-SNE to test the effect. When there are few features, it can be seen from the figure that only 2 small clusters are not distributed according to the ideal.

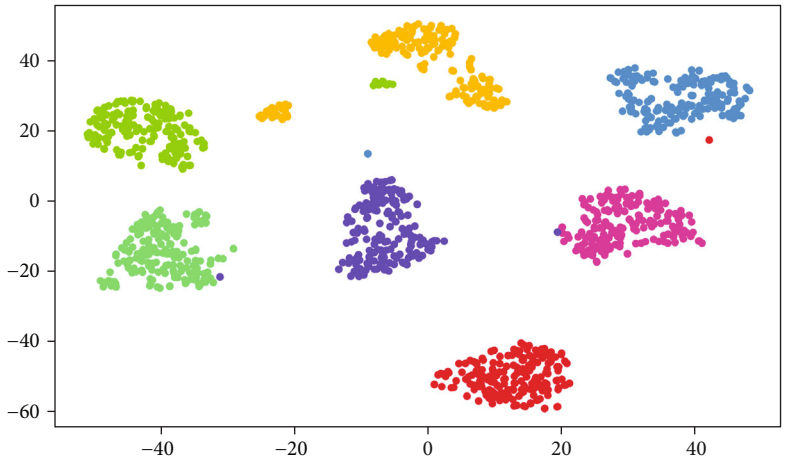


Figure 2: t-SNE for 0-6 digits. When the features increase, the data reduced by t-SNE is more difficult to be analyzed by clustering.

solve this problem by using t-distribution instead of Gaussian distribution. Equation (6) shows the result after replacing with t-distribution.

$$q_{j|i} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}. \tag{6}$$

KL distance is to find the optimal value.

$$\frac{\delta C}{\delta y_i} = 4\sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}. \tag{7}$$

However, we find that the FDP algorithm cannot effectively cluster the data after t-SNE dimensionality reduction.

In the process of Figures 1–3, the data in Figure 1 only has six groups of features. After dimensionality reduction, we can see that each class is separated and not mixed together. Figure 2 has 7 groups of features. At this time, a small green piece appears near a large yellow block, and a small yellow one appears near the green one. If a clustering algorithm is used, the green class will be divided into yellow, and the yellow small piece is divided into green, which leads to the final result error of clustering. In this picture, the same is true. There are many other cases which will lead to the error of the final clustering results. But there are 10 groups of data in Figure 3. It should have appeared in the same region with the same color. However, it can be seen from the graph that many colors are mixed together and are not effectively separated. If clustering algorithm is used at this time, the accuracy of clustering algorithm will be reduced, and even the wrong results will be obtained. In the next section, we propose an improvement.

3. Algorithm Improvement

3.1. Feature Extraction. Figures 1–3 show the process in which the effect of the algorithm becomes worse as the feature increases. t-SNE [19] has decreased with the increase of data features. The cluster algorithm cannot be correctly

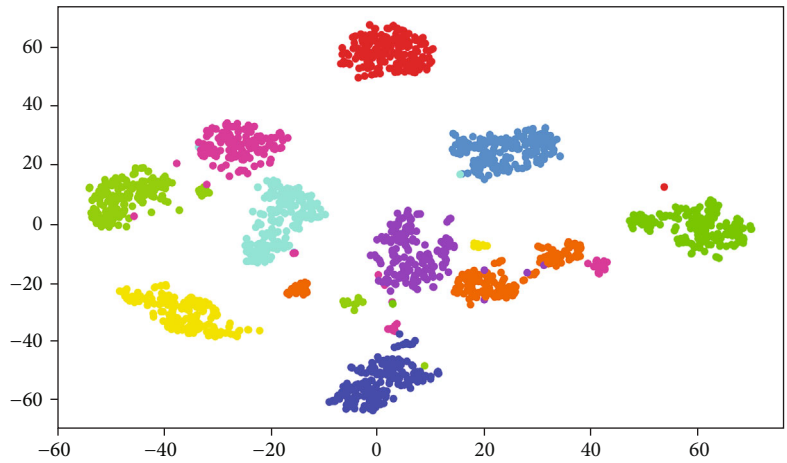


Figure 3: t-SNE for 0-9 digits. When the features are further increased, the distribution of data is more unclear, which is not conducive to clustering algorithm.

distributed to the right location. This will inevitably degrade the performance of cluster algorithm. We decided to introduce random forest to improve the performance.

The performance of traditional t-SNE still has some problems, so we decided to introduce random forest to improve the performance. The data in scientific research and production life are very complex, but the main features of these data that can distinguish each other are sometimes not many. For example, fruit can be distinguished by shape and color. Therefore, effectively selecting the main features can improve the performance of the t-SNE algorithm. The feature extraction of random forest [20] is mainly based on the out of bag (OOB) principle. If a feature is important, then when a certain amount of noise is introduced into the distributed data of this feature, the performance of the model should be greatly changed by random forest training with only the changed data of this feature. On the contrary, if a feature is unimportant, the performance of the retrained model will not change much. Data has many features. We need to calculate the importance value of all features. Firstly, a random forest is established, and the decision tree of the random forest uses C4.5. Firstly, for a feature, the error of its packet data is calculated according to the established random forest, which is recorded as errOOB1. Then, the interference data is added to the same feature to calculate errOOB2.

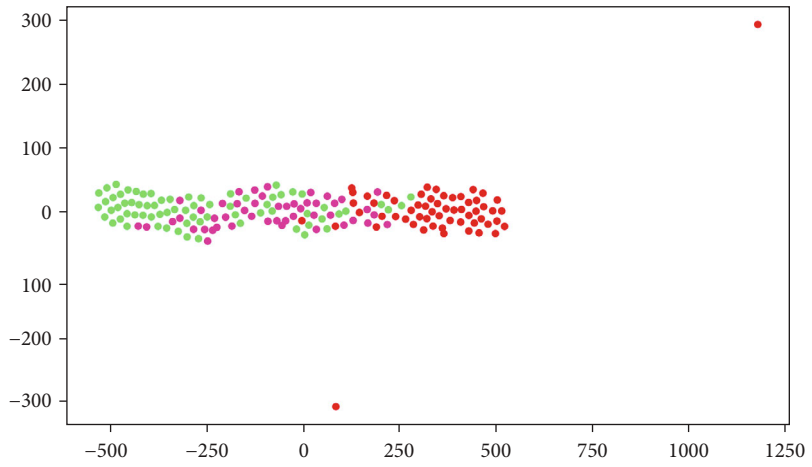
$$\text{Importance}_x = \frac{\sum_{i=1}^{N}(\text{errOOB2} - \text{errOOB1})}{N}. \quad (8)$$

N stands for n trees. If we add noise to a feature at random, Importance_x will be greatly reduced, which means that the main feature. It has a high degree of importance. The feature select process is to sort the feature variables by random forest in descending order of Importance_x. A new feature set is obtained by determining the deletion ratio and eliminating the unimportant indexes from the current feature variables. After calculating the importance value of all the features, we calculate their average value. If the importance value is less than 10%-20% of the mean value, the remaining features are the main features. After dimensionality reduction of these features, the effect of FDP will increase. Figures 4 and 5 show the effect after extracting the main features. t-SNE will further put the original class together instead of leaving a class. t-SNE PCA and locally linear embedding [21] (LLE) cannot effectively separate the data, which affects the clustering effect.

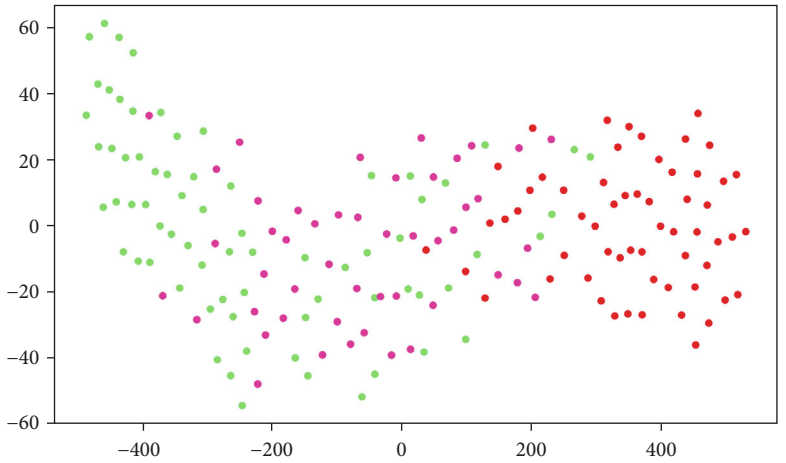
1: Input: $D = \{x_1, x_2, \cdots, x_m\}^n$.
2: Output: $Y = \{y_1, y_2, \cdots, y_m\}^2$.
3: Build random forest.
4: Compute Importance_x.
5: Compute avg(Importance_x).
6: for 1 to n.
7: if $\text{Importance}_i <= \text{avg}$.
8: Remove feature.
9: end.
10: end.
11: Generate $G = \{x_1, x_2, \cdots, x_m\}^r$.
12: G use by t-SNE.
13: Generate $Y = \{y_1, y_2, \cdots, y_m\}^2$.

Algorithm 1: The improvement of t-SNE.

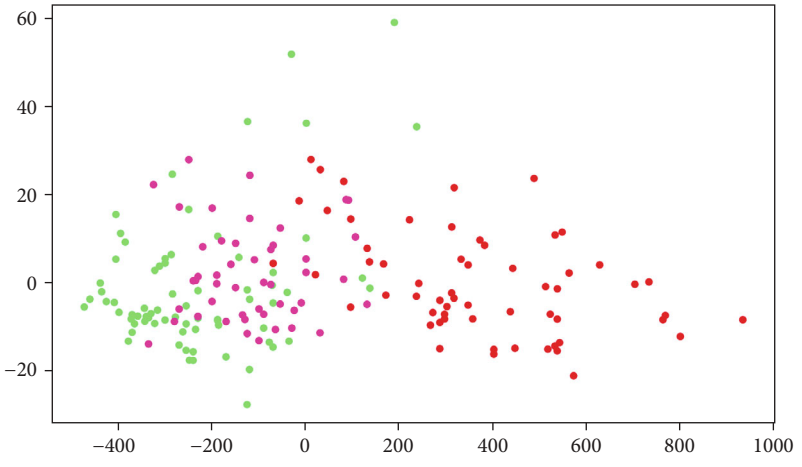
Now, let us take a look at Figure 4. This picture shows the comparison between the improved t-SNE and t-SNE and PCA and LLE. In Figure 4, the improved t-SNE, t-SNE, PCA, and LLE are represented by (a), (b), (c), and (d), respectively. We use the wine dataset in the UCI dataset, which has 3 categories and 13 features. The goal of this experiment is to reduce this set of data to 2 dimensions. We use the same color to represent the same class and use four algorithms to reduce the original data to 2 dimensions. It can be seen that t-SNE, PCA, and LLE do not effectively divide the same color into the same area. In t-SNE, in the -200 to 200 regions, three colors are mixed together. In PCA, in the -200 to 200 regions, the three colors are mixed together, so is LLE. In the improved t-SNE, green, purple, and blue are effectively



(a)



(b)



(c)

Figure 4: Continued.

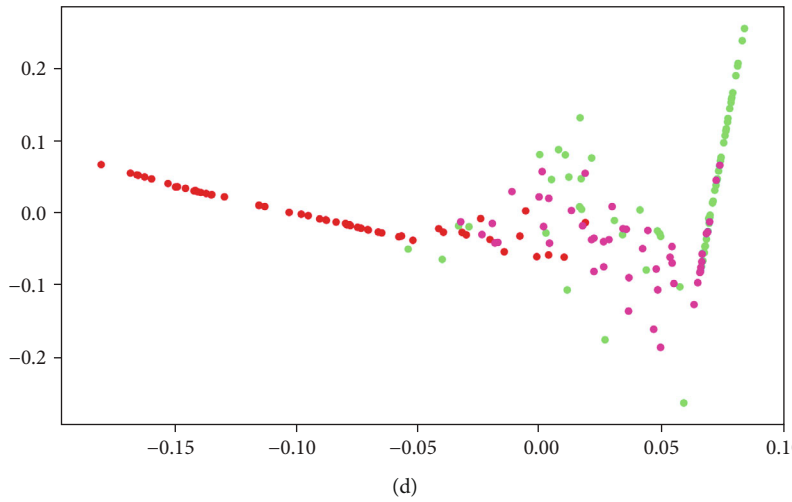


Figure 4: Comparison chart in wine datasets. (a) Improved t-SNE. (b) t-SNE. (c) PCA. (d) LLE.

divided into three parts. In this way, the clustering algorithm can cluster more effectively.

In Figure 5, the improved t-SNE, t-SNE, PCA, and LLE are represented by (a), (b), (c) and (d), respectively. We use the digits dataset in the UCI dataset, which is 64-dimensional data. The goal of this experiment is to reduce this set of data to 2 dimensions. We use the same color to represent the same class and use four algorithms to reduce the original data to 2 dimensions. First of all, you can see that all the colors in c are mixed up in a disorderly way. After dimensionality reduction, each class is not separated effectively. This will lead to the subsequent clustering effect worse or even get the wrong result. In (d), although each color is roughly in a straight line, except for fans and dark blue which are obviously separated, other colors are mixed together and not effectively separated. The effect in (b) is better than that in (c) and (d), but we can see that a large number of small pieces are not assigned to the corresponding position. For example, a small knob of light blue is far away from the original a lumpen mass of light blue, and this small knob of light blue is closer to other colors, which will lead to the subsequent clustering algorithm cannot effectively classify. It will lead to the result error. The improved algorithm can reduce the occurrence of two cases, so it can improve the accuracy of clustering algorithm after dimension reduction.

3.2. Self-Adaptive Selection. Now, let us talk about FDP. In our further experiments, we found a lot of problems and defects about the algorithm itself. The selection of d_c and the judgment of the decision graph will affect the result of the final clustering algorithm. In this paper, we introduce and improve the decision graph selection problem. We will further improve the selection of d_c in the follow-up work. The improvement of d_c selection is also very meaningful. Back to the part of the decision graph, we find that the experimenter needs to decide the cluster center by himself. This greatly affects the clustering performance and effect. When artificial selection is introduced into the algorithm, the accuracy of the algorithm will be greatly reduced. If this algorithm needs to be promoted and expanded, its ease of use will be greatly reduced. Through a large number of experiments, we solve the problem of manually selecting the center point and improve the accuracy of the original algorithm.

For the problem that the center point cannot be selected effectively in FDP, we design a new judgment system through the difference and change rate, so that it can select the center adaptively. It does not need to check manually to complete the center point selection. We should make full use of ρ_i and δ_i. Firstly, ρ_i and δ_i are multiplied, and the value is set as λ_i

$$\lambda_i = \rho_i \cdot \delta_i, \tag{9}$$

$$\Delta_i = \lambda_i - \lambda_{i+1}, \tag{10}$$

$$\theta_i = \frac{\max(\Delta_i, \Delta_{i+1})}{\min(\Delta_i, \Delta_{i+1})}. \tag{11}$$

The improved algorithm relies on the calculation of the λ. We sort and start with 1. λ_1 is biggest, and so on. And then, we make use of mathematical analysis of the results λ. Firstly, we use Equation (10) to get the difference. The main idea of formula (10) is to make difference Δ between two adjacent terms. Then, the calculated difference Δ is processed by Equation (11). The main idea of Equation (11) is to get the change rate of adjacent difference θ. These two values are regarded as the prerequisite of the core point judgment.

This is the pseudocode described by the algorithm RCD-FDP.

At this time, we calculate the arithmetic mean value of Δ and θ to get avg(Δ) and avg(θ), respectively. At this point, we can transform the problem into comparing Δ and θ with avg(Δ) and avg(θ). After a series of experiments, when Δ is less than 10%-25% of the avg(Δ) and θ is less than 50% of the avg(θ). You can stop judging and get the center point. Table 1 is a manual dataset. Due to the large amount of data, only the first 11 sample data are shown in this paper. It can be seen that if we judge by hand, we cannot accurately judge the cluster center. However, by introducing Δ, θ, avg(Δ), and avg(θ), the core point can be determined adaptively. It avoids

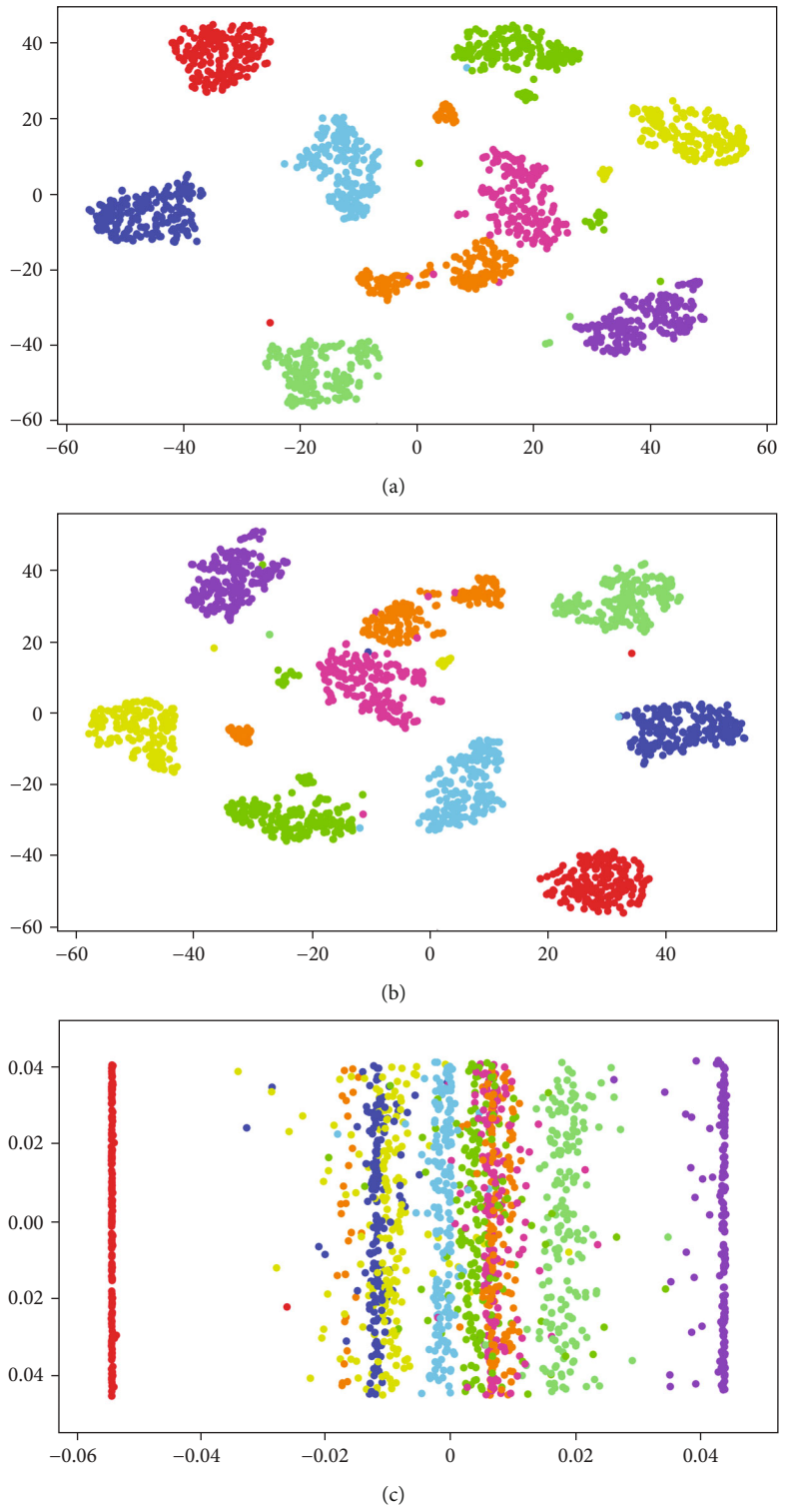
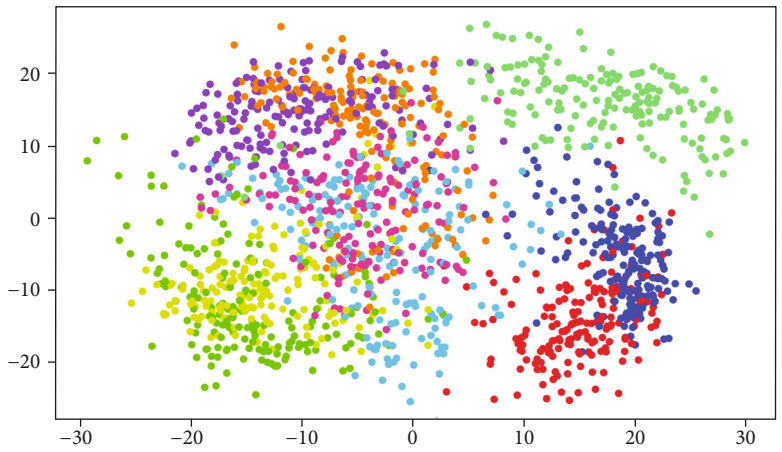


Figure 5: Continued.



(d)

Figure 5: Comparison chart in digits datasets. (a) Improved t-SNE. (b) t-SNE. (c) PCA. (d) LLE.

1: Input: $D = \{x_1, x_2, \cdots, x_m\}^n$, Δ, θ.
2: Output: cluster class y_i belong to $x_i \in$ D.
3: Compute and sort λ_i.
4: **for** 1 to m.
5: $\Delta_i = \lambda_i - \lambda_{i+1}$.
6: end.
7: **for** 1 to m.
8: $a = \max(\Delta_i, \Delta_{i+1})$.
9: $b = \min(\Delta_i, \Delta_{i+1})$.
10: $\theta_i = a/b$.
11: end.
12: **for** 1 to m.
13: if $\Delta_i < \Delta$ && $\theta_i < \theta$.
14: cluster number = $i - 1$.
15: end.
16: Cluster by FDP.

Algorithm 2: RCD-FDP.

the ground error caused by manual judgment. The above pseudocode RCD-FDP shows the process of the algorithm.

Combining the data after improved t-SNE dimension reduction with RCD-FDP, this algorithm is called IT-RCD-FDP, which has an obvious effect on high-dimensional datasets.

4. Experiment

4.1. Evaluation Criterion. This section shows the performance of the improved algorithm through experimental comparison. In this experiment, digits, wine, and heart disease in the UCI dataset are selected. Through these datasets, the experiment uses some commonly used evaluation criteria [22] to compare with the original FDP algorithm, K-means algorithm, DBSCAN, and spectral clustering. Four evaluation criteria are used in this experiment. It includes accuracy, adjusting rand index, normalized mutual information, and completeness.

Table 1: Change rate and difference.

i	ρ	δ	Δ	θ
1	17.1	21.15	240.455	12.4
2	10.91	11.11	19.38	4
3	8.53	11.95	79.055	8.7
4	6.1	3.75	8.995	1.3
5	4.13	3.36	6.46	3.8
6	5.26	1.41	1.668	8.9
7	7.19	0.8	0.186	2.1
8	5.06	1.1	0.406	2.3
9	1.66	3.11	0.94	1.2
10	16.88	0.25	1.125	10
11	6.19	0.5	0.105	1.7

Accuracy [23] is one of the most commonly used means, which is often used in binary classification. In clustering, we only need to transform multiclassification into a two-classification problem by transforming the problem into a judgment of consistency. If the tested algorithm used is very good, the calculated value is 1. The accuracy of the evaluated algorithm can be calculated by (12).

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{12}$$

Rand index [24] is judged and calculated by comparing tags with existing results. It calculates the number of the same cluster and the number of different clusters by judging whether it is a cluster. But RI has problems with random tags. At this time, adjusting RI [25] was proposed to solve this problem. Equations (13) and (14) show the calculation process of RI and ARI

$$\mathrm{RI} = \frac{a + d}{a + b + c + d}, \tag{13}$$

$$\mathrm{ARI} = \frac{\mathrm{RI} - E(\mathrm{RI})}{\max(\mathrm{RI}) - E(\mathrm{RI})}. \tag{14}$$

In view of the knowledge of the assignment of base truth classes and our prediction of the assignment of clustering algorithms for the same sample, mutual information [26] (MI) is the degree of closeness between the two is expressed in the form of joint probability. It can measure the clustering result of the clustering algorithm. It compares the joint probability density of the existing tags and the clustering algorithms. Normalized MI [27] is obtained by calculating the arithmetic mean of MI. Equations (15), (16), and (17) show the calculation process of MI and NMI.

$$H(X) = -\sum_i p(x_i) \log(x_i), \tag{15}$$

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{16}$$

$$\mathrm{NMI}(X;Y) = 2\frac{I(X;Y)}{H(X) + H(Y)}. \tag{17}$$

Completeness [28, 29] is a measure of cluster labels given ground truth. Its idea is illustrated by an example: there are three classes, and each class has a fixed arrangement of students. When all the students go back to their class, its value is the largest. But when the clustering algorithm guides the students back to the class, there will be errors, so that the students who should not be in this class will appear in this class. Then, the value will decrease.

4.2. Data and Comparison Experiments. The experimental datasets are the digits dataset, wine dataset, and heart disease (Cleveland) dataset. Through following UCI dataset experiments, we can see that it a has good performance under the four evaluation indexes. The UCI dataset wine is 13-dimensional, UCI dataset heart disease is 14-dimensional, and UCI dataset digits is 64-dimensional. Experiments are compared with the original FDP, *K*-means, DBSCAN, and spectral clustering algorithm, through the above evaluation criteria.

For each set of datasets, we use the idea of 10-fold cross-validation. Each dataset is divided into ten parts according to the number of samples and labeled from 1 to 10. In the first experiment, the part marked 1 is removed and the remaining data is used for testing. At this time, the values of all clustering algorithms under different evaluation indexes are obtained. According to this method, 10 sets of data are obtained. For these 10 groups of data, for example, ACC of the IT-RCD-FDP part has 10 results. We remove the maximum and minimum values and then calculate the draw value as the final experimental result. For *K*-means algorithm, the center point needs to be randomly selected during initialization. This brings a certain amount of randomness. In order to solve this problem, we select the first k sample values as the center in the experiment. In this way, the randomness can be eliminated. After all the data are processed above, results are obtained in Tables 2–4.

Table 2: UCI experiments digits.

Type	ACC	NMI	ARI	Completeness
FDP	0.51	0.56	0.53	0.41
K-means	0.49	0.73	0.66	0.74
DBSCAN	0.59	**0.75**	0.63	0.66
Spectral	0.53	0.61	0.67	0.63
IT-RCD-FDP	**0.66**	0.71	**0.77**	**0.78**

Table 3: UCI experiments wine.

Type	ACC	NMI	ARI	Completeness
FDP	0.51	0.54	0.68	0.64
K-means	0.35	0.42	0.37	0.38
DBSCAN	0.61	0.51	0.65	0.63
Spectral	0.62	**0.73**	0.71	0.61
IT-RCD-FDP	**0.64**	0.61	**0.73**	**0.66**

Table 4: UCI experiments heart disease (Cleveland).

Type	ACC	NMI	ARI	Completeness
FDP	0.09	0.17	0.08	0.06
K-means	0.12	0.19	0.14	0.12
DBSCAN	0.01	0.03	0.03	0.05
Spectral	**0.15**	0.16	0.11	0.12
IT-RCD-FDP	0.13	**0.21**	**0.25**	**0.23**

Table 2 shows a set of comparison results of four different evaluations using UCI dataset digits, among which IT-RCD-FDP is our improved algorithm. The results show that ACC, ARI, and completeness are the best. In the digits dataset, the evaluation index NMI of IT-RCD-FDP is 7.2% higher than other algorithms. Table 3 also shows a set of comparison results of four different evaluations using UCI dataset wine. The results show that ACC, ARI, and completeness are the best. In the wine dataset, the evaluation index NMI of IT-RCD-FDP is 11% higher than other algorithms. According to wine and digits, the evaluation index NMI of IT-RCD-FDP is 23% higher than that of the original FDP. As can be seen from Table 4, although IT-RCD-FDP scored the highest on NMI, ARI, and completeness. But the overall level is very low. This also shows that the clustering algorithm is weak in the medical field. We hope to break through these scenes in the later work.

4.3. Application. Now, let us talk about applications. As mentioned above, clustering algorithm also has good applications in wireless sensor networks. Because the wireless sensors distributed in the space are just like every data point in the dataset. They all have their own characteristics and attributes. How to manage and control these sensors needs to understand them and analyze them. If wireless sensors are regarded as data points, clustering can be used to study and analyze them. In wireless sensor networks, the selection of the cluster head is particularly important. The cluster head is just like

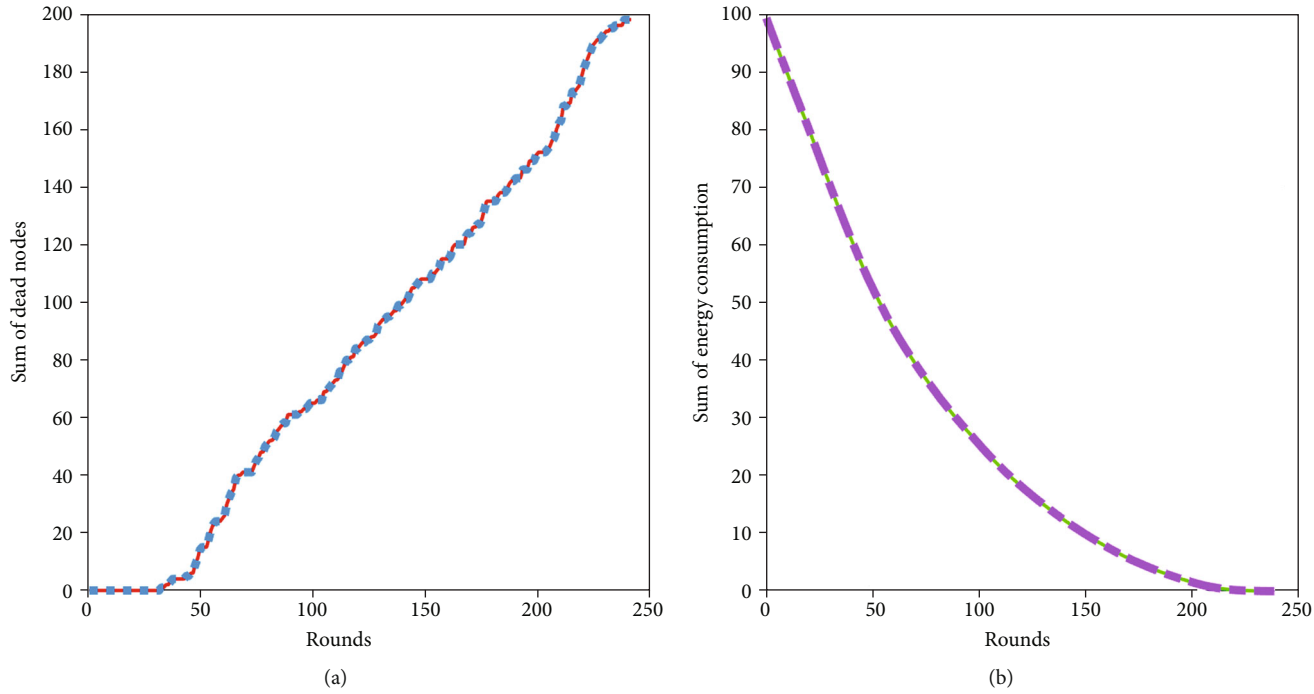


Figure 6: Performance diagram. Clustering algorithm is applied to wireless sensor network performance demonstration.

the cluster center. In this way, the cluster head selection problem is transformed into the cluster center selection problem. The algorithm will cluster according to the centers. In many algorithms, low-energy adaptive clustering hierarchy (LEACH) algorithm [30] has the most reliable and efficient. Although LEACH performance is very good, there is also the problem that the selected cluster heads are too concentrated. To solve this problem, we introduce IT-RCD-FDP into LEACH to solve this problem. IT-RCD-FDP algorithm is used to cluster all the sensors in a wireless sensor network. Due to the adaptive nature of it, cluster heads can be automatically selected and clustered according to the cluster heads. In this way, wireless sensor networks are automatically divided into several clusters. Then, a sensor with the largest energy is selected as the cluster head in each round of the cluster. In this way, we can solve the problems of LEACH. Figure 6 shows the performance and effect of applying IT-RCD-FDP to LEACH.

5. Conclusions

This paper starts from two problems of FDP. Firstly, FDP algorithm is difficult to deal with high-dimensional data. This paper introduces the improved t-SNE algorithm. After such processing, the ability of the original algorithm to process high-dimensional data is improved. Secondly, FDP algorithm cannot select centers adaptively. In this paper, the change rate and difference are introduced to make the original algorithm select the centers adaptively. Finally, we apply the improved algorithm in WSN cluster head selection and achieve good results. The IT-RCD-FDP proposes that new way to solve high-dimensional data. Compared with the original algorithm, the algorithm finds a threshold point by a mathematical method and improves the original manual judgment method to automatic operation by setting the range. In this way, the accuracy of the algorithm can be greatly improved. Through the improvement of t-SNE, the result of FDP is more accurate.

Data Availability

Previously reported data were used to support this study and are available at url = "http://archive.ics.uci.edu/ml." These prior studies are cited at relevant places within the text as references.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

First of all, I would like to express my gratitude to my parents for sheltering me from the wind and rain so that I can thrive. Secondly, I would like to thank Northwest Normal University. I would like to express my gratitude to my tutor for teaching and helping. Finally, I would like to express my gratitude to my classmates for lighting up my mind, when I am not happy and upset. When I encountered a bottleneck in my writing, the discussion and analysis with them gave me infinite inspiration. When I am restless, they are at my side to help me analyze problems and provide feasible suggestions. This work was supported by the Northwest Normal University under Grant (Research on Retina Image Segmentation and Aided Diagnosis Technology based on Deep Learning) 61962054.

References

[1] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: a K-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100–108, 1979.

[2] Y. J. Zhang, T. Oka, R. Suzuki, J. T. Ye, and Y. Iwasa, "Electrically switchable chiral light-emitting transistor," *Science*, vol. 344, no. 6185, pp. 725–728, 2014.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, vol. 8, pp. 226–231, Miinchen, German, 1996.

[4] E. Nasibov and G. Ulutagay, "Robustness of density-based clustering methods with various neighborhood relations," *Fuzzy Sets and Systems*, vol. 160, no. 24, pp. 3601–3615, 2009.

[5] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[6] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[7] C. Kourouklas, O. Mangira, A. Iliopoulos, D. Chorozoglou, and E. Papadimitriou, "A study of short-term spatiotemporal clustering features of Greek seismicity," *Journal of Seismology*, vol. 24, no. 3, pp. 459–477, 2020.

[8] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[9] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[10] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Systems with Applications*, vol. 150, p. 113277, 2020.

[11] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[12] X.-D. Wang, R.-C. Chen, Z.-Q. Zeng, C.-Q. Hong, and F. Yan, "Robust dimension reduction for clustering with local adaptive learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 657–669, 2019.

[13] S. Wang and Y. Li, "A dynamic cluster model based on projection pursuit with its application to climate zoning," *Journal of Applied Meteorological Science*, vol. 18, no. 5, pp. 722–726, 2007.

[14] 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, 2016.

[15] X. Xue, X. Wu, C. Jiang, G. Mao, and H. Zhu, "Integrating sensor ontologies with global and local alignment extractions," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6625184, 10 pages, 2021.

[16] 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.

[17] X. Xue, C. Yang, C. Jiang, P.-W. Tsai, G. Mao, and H. Zhu, "Optimizing ontology alignment through linkage learning on entity correspondences," *Complexity*, vol. 2021, Article ID 5574732, 12 pages, 2021.

[18] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," *Advances in Neural Information Processing Systems*, vol. 15, pp. 833–840, 2003.

[19] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] D. Liu and K. Sun, "Random forest solar power forecast based on classification optimization," *Energy*, vol. 187, p. 115940, 2019.

[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[22] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[23] J. S. Smith, B. T. Nebgen, R. Zubatyuk et al., "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning," *Nature Communications*, vol. 10, no. 1, article 2903, 2019.

[24] R. J. G. B. Campello, "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[25] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Artificial Neural Networks – ICANN 2009*, vol. 5769, pp. 175–184, Springer, 2009.

[26] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[27] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[28] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification," *Fuzzy Sets and Systems*, vol. 155, no. 2, pp. 191–214, 2005.

[29] O. Arbelaitz, I. F. Gurrutxaga, J. Muguerza, J. M. Perez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[30] K. Lutful, N. Nidal, and T. Sheltami, "A fault-tolerant energy-efficient clustering protocol of a wireless sensor network," *Wireless Communications and Mobile Computing*, vol. 14, no. 2, pp. 175–185, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6649127, 11 pages
https://doi.org/10.1155/2021/6649127

Research Article

Research on Security Level Evaluation Method for Cascading Trips Based on WSN

Hui-Qiong Deng,[1,2] Jie Luo,[1,2] Kuo-Chi Chang,[3,4] Qin-Bin Li,[1] Rong-Jin Zheng,[1] and Pei-Qiang Li[1]

[1]*School of Electronic Electrical and Physics, Fujian University of Technology, Fuzhou 350118, China*
[2]*Fujian Provincial University Engineering Research Center for Simulation Analysis and Integrated Control of Smart Grid, Fujian, China*
[3]*College of Mechanical & Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan*
[4]*Department of Business Administration, North Borneo University College, Sabah 88400, Malaysia*

Correspondence should be addressed to Jie Luo; 550461238@qq.com

Received 28 December 2020; Revised 8 March 2021; Accepted 26 April 2021; Published 19 May 2021

Academic Editor: Pei-Wei Tsai

In recent years, the application of wireless sensor networks (WSN) in power systems has received a great deal of attention. As we all know, the most important issue for the power system is security and stability, especially due to the massive outages caused by cascading trips. Therefore, in today's era, from the perspective of cascading trips, how to effectively use WSN to analyze and evaluate the security level of the power grid is an important direction for future power development. In this paper, an algorithm based on the WSN collection of online data to calculate the corresponding security level of the system is proposed for the cascading trip phenomenon, to achieve the online evaluation of the cascading trips. First, this paper proposes a hybrid layered network structure based on WSN for monitoring system and details the acquisition of power grid parameters by its acquisition layer. Secondly, combined with the manifestation of cascading trips and the action equation of current-type line backup protection, the mathematical representation of the grid cascading trips is given, and the mathematical form corresponding to the critical situation is strictly proved, and an index for evaluating the security level of the power grid is proposed and then further combined with the actual physical constraints of the power grid and the establishment of a mathematical model for calculating the security level of the grid cascading trips. For this model, this paper relies on evolution particle swarm optimization (EPSO) to give specific ideas for solving the model. Finally, a case analysis is performed by the IEEE39 node system and the results of the case show the effectiveness of the model and method.

1. Introduction

With the development of microsensor technology, microelectronics technology, wireless communication technology, and computer technology, wireless sensor networks (WSN) with the functions of information collection, processing, and transmission emerge as the times require [1]. Currently, WSN technology is also gradually penetrating into the power industry, and its application prospects have attracted much attention.

As a system that provides clean secondary energy to cities and villages, today's power system has established a very strong connection to society as a whole. Therefore, the safe and reliable operation of the power system becomes an essential topic, and to ensure the safe and reliable operation of the power system, first of all, we need to quickly and accurately collect the power system operation data. Then, analyze on this basis to determine whether the settings of the various operating parameters of the power system are reasonable, whether they need to be adjusted, and how to adjust them. Conventional power systems typically use potential transformers and current transformers to collect power data, gather them to the monitoring center of the substation, and then transmit the data to a remote dispatch center with

corresponding communication facilities. The advantages of this method are high security, stability, the anti-interference ability of data transmission, small size, and lightweight. However, this method also has some obvious disadvantages, such as the existence of certain limitations, comprehensive wiring difficulties, long construction period, high cost, limited monitoring range, poor scalability, equipment maintenance difficulties, and a series of other problems [2].

WSN can efficiently and quickly collect and transmit the main data of the whole system of the power system due to its low cost, adaptability to the environment, efficient collection of information, and full coverage of the monitoring area [3]. WSN has emerged as the ideal choice to meet the new challenges of power grid parameter monitoring technology. The wireless sensor installed on the power equipment is used to complete electrical information acquisition and preprocessing., and the synchronized data collected will be transmitted to the monitoring center through the wireless communication network, which analyzes and processes the information [4]. WSN for power grid parameters to provide more flexible and complete monitoring solutions can deal with the power grid development and application of special requirements and can realize centralized management of multitype power grid parameter monitoring [5]. The above features of WSN for the overall perspective of the power system security analysis and defense are extremely advantageous, especially for the analysis and prevention of cascading trips.

Generally speaking, the cascading trips of the power system is an event caused by the interaction between components, which in severe cases can lead to vicious blackouts, so cascading trips have attracted much attention. At present, researchers have done a great deal of work in the field of defense and control regarding cascading trips. The literature [6] addresses the problem of cascading trips, by predicting the state of the monitoring node set in the process of cascading trips; traction control is implemented for nodes in abnormal states to inhibit the propagation of failure in the network. However, it is difficult to apply it in practice due to insufficient consideration of the operating characteristics of the grid. The literature [7] proposes a wide-area collaborative precontrol method based on the theory of multi-intelligent systems and the analysis of offline cross-sectional power transmission limits. However, the influence of the current operation state on the outage probability of power system components is ignored. The literature [8] developed a model of cascading trip network interaction with information network edges as initial faults and proposed a vector construction method for false data attacks based on parameter estimation. Finally, corresponding defense measures are proposed based on false data attacks. However, the dynamic characteristics of the power grid are not considered. The literature [9] proposes a cascading trip path search and warning model based on the characteristics of the power grid information physical fusion system. However, there is the problem of inaccurate system data.

Given the actual characteristics of WSN and power grid cascading trips, this paper proposes a method to measure nodal power injection using WSN and analyze the security of the power grid for the expected initial failure to trigger cascading failure. The second and third parts of this paper mainly introduce the application of WSN in power system and the monitoring system based on WSN; the fourth part gives the mathematical equation to judge the cascading trips of the power grid; the fifth part mainly introduces how to use the node injection power data collected by WSN to analyze the security level of the power grid for cascading failure, and gives the specific analysis model; the sixth part mainly gives the solution algorithm and process for the given model; the seventh part gives an example based on the IEEE39 node system to verify and analyze the method in this paper.

2. Combination of WSN and Power Grid Monitoring

WSN is the core of the Internet of Things technology. It is one of the most cutting-edge technologies to realize the acquisition and transmission of various signals through low-power self-organizing and adaptive wireless sensor nodes [10]. The traditional wired communication wiring is cumbersome, the line is easy to aging, and the cost is high, while the wireless sensor network fully meets the speed requirements of power equipment condition monitoring, perfectly solves these shortcomings, reduces the cost of power operation and maintenance costs, and improves the stability and efficiency of power system operation. It promotes the application and rapid development of wireless sensor network technology in the power grid and lays a solid foundation for the efficient and rapid construction of power system network framework in the future and the improvement of user satisfaction.

.In the power system, the use of WSN technology to establish a remote monitoring system, at any time to monitor the status of power equipment data, to help operators on-line security assessment of the status of power equipment, abnormal response to the characteristics of the quantity, to take the necessary measures to avoid the occurrence of serious failure. In the distribution network relay protection, the wireless current sensor using WSN technology not only solves the problem of possible subcurrent saturation of the current transformer but also is easy to install, while the current data of the line is accurately and quickly collected by WSN to avoid the initial fault triggering the current overload protection and causing the relay device to operate incorrectly [11]. In summary, the combination of WSN and grid monitoring is a new trend in the future development of smart power grids.

3. Monitoring System Based on WSN

At present, the power system monitoring network based on WSN is mainly linear distribution, which usually arranges sensor nodes on transmission lines and relay nodes on towers and sends the collected information through the sensor nodes of the lines to the relay nodes of the towers, which are processed by the relay nodes and forwarded to the substations. This monitoring network has disadvantages such as poor effectiveness and can cause uneven load distribution. The use of a layered network architecture avoids the

problems of high cost, low reliability, poor scalability, and limited transmission rate brought about by the existing monitoring systems that rely entirely on mobile communication networks. Therefore, in the development of smart power grids, the introduction of the WSN, combined with fiber optic Ethernet, can make full use of existing resources and protect existing investment, which is a more suitable choice in the current situation.

3.1. Structure of the Monitoring Network. This paper proposes a hybrid layered network structure to monitor the power system network by using the methods of literature [12, 13], combining the wired (optical fiber) and wireless (ZigBee and cellular) technology. The structure of the network is shown in Figure 1.

The whole network structure is divided into the acquisition layer, convergence layer, and teleportation layer. The acquisition layer is composed of a large number of sensor nodes, which is responsible for collecting electrical information on the transmission line, and a relay node with strong processing ability is arranged on each tower. The convergence layer consists of these relay nodes that are responsible for receiving and processing ordinary node data on the pole and tower. The teleportation layer is composed of representative nodes, substations, and monitoring centers, with wireless ZigBee connections between relay nodes and between relay nodes and substations, and data transmission from substations and representative nodes to monitoring centers via optical fiber and cellular networks, respectively. This hierarchical structure has the characteristics of strong network extensibility, strong effectiveness, and easy centralized management, which can meet the application requirements of the emerging smart power grids.

3.2. Acquisition of the Power Grid Parameters. The power grid parameters, which serve as accurate indexes for evaluation and feedback on the operational status of the power grid, must be dynamically monitored in real time. In this paper, with the help of new sensor technology, fast, accurate, and comprehensive realization of wireless acquisition of power grid parameters. Therefore, this paper focuses on the acquisition layer of the monitoring network structure. The hardware system design of the acquisition layer is shown in Figure 2.

As shown in Figure 2, in the acquisition layer, the wireless sensor acquisition nodes deployed in the power grid are responsible for voltage and current acquisition and processing and wireless transmission. It mainly includes a data acquisition module, a wireless transceiver circuit, and a power supply circuit. The power grid parameter acquisition node collects current signals and voltage signals through high-precision current transformers and voltage transformers and then processes them through analog conditioning circuits and inputs them to the sampling unit of CC2530 to complete the A/D conversion of the signals. Finally, wireless communication is realized by an RF transceiver. The power module is powered by two dry batteries.

The voltage phasor of the node and the current phasor of all the branches connected to it are obtained by converting the measured sample values into frequency domain signals:

$$\begin{cases} \dot{\mathbf{U}} = \dfrac{2}{N}\sum_{k=0}^{N-1} u_k e^{-j2\pi/Nk} = \mathbf{U_r} + j\mathbf{U}_\mathrm{d}, \\ \dot{\mathbf{I}} = \dfrac{2}{N}\sum_{k=0}^{N-1} i_k e^{-j2\pi/Nk} = \mathbf{I}_\mathrm{r} + j\mathbf{I}_\mathrm{d}, \end{cases} \tag{1}$$

where u_k and i_k denote the kth voltage sampling value and current sampling value, respectively; N is the total number of samples; $\mathbf{U}_\mathrm{r}$ and $\mathbf{I}_\mathrm{r}$ denote the real part of the phasor; and $\mathbf{U}_\mathrm{d}$ and $\mathbf{I}_\mathrm{d}$ denote the imaginary part of the phasor.

Based on the voltage phasor and current phasor obtained above, the nodal injection power can be further calculated using these data, and its calculation formula is as follows:

$$\begin{cases} S_{ij} = \dot{U}_i \times \left(\dot{I}_{ij}\right)^*, \\ S_i = \sum_{j\in \mathrm{i}} S_{ij}, \end{cases} \tag{2}$$

where $U_i^{\cdot}$ is the voltage phasor of node i; $\dot{I}_{ij}$ is the current flowing through branch road L_{ij}; S_{ij} is the flow power on the i-side of the branch road L_{ij}; and S_i is the injection power of node i.

4. Basic Idea for Evaluating Power Grid Security Level Based on WSN

As mentioned above, the WSN is used to get the power injection data of the power grid nodes, and then, the data is transmitted to the monitoring center through a hierarchical structure. This chapter will introduce how to use the nodal injection power data to help the monitoring personnel to evaluate the safety level of the cascading trips, and give the relevant expression and mathematical model.

4.1. Mathematical Representation of Power Grid for Cascading Trips. The cascading trips of the power grid is usually after the initial failure branch road is removed; due to the redistribution of power flow, the backup protection action in the branch road except the initial failure branch road is caused.The cascading trips of the power grid is usually caused by the action of backup protection in branch roads other than the initial failure branch road due to the redistribution of power flow after the removal of the initial failure branch road. Taking the current protection as an example, if the branch road L_a in a power grid has an initial failure at a certain time, then whether any branch road L_b in the remaining system in the power grid will have a cascading trip after the branch road L_a is cut off; it can be judged by whether the current detected by its configured backup protection enters the action zone of the protection. The branch road L_b is between node i and node j. If the node i-side of the branch road L_b is equipped with current-type backup protection, the equation shown

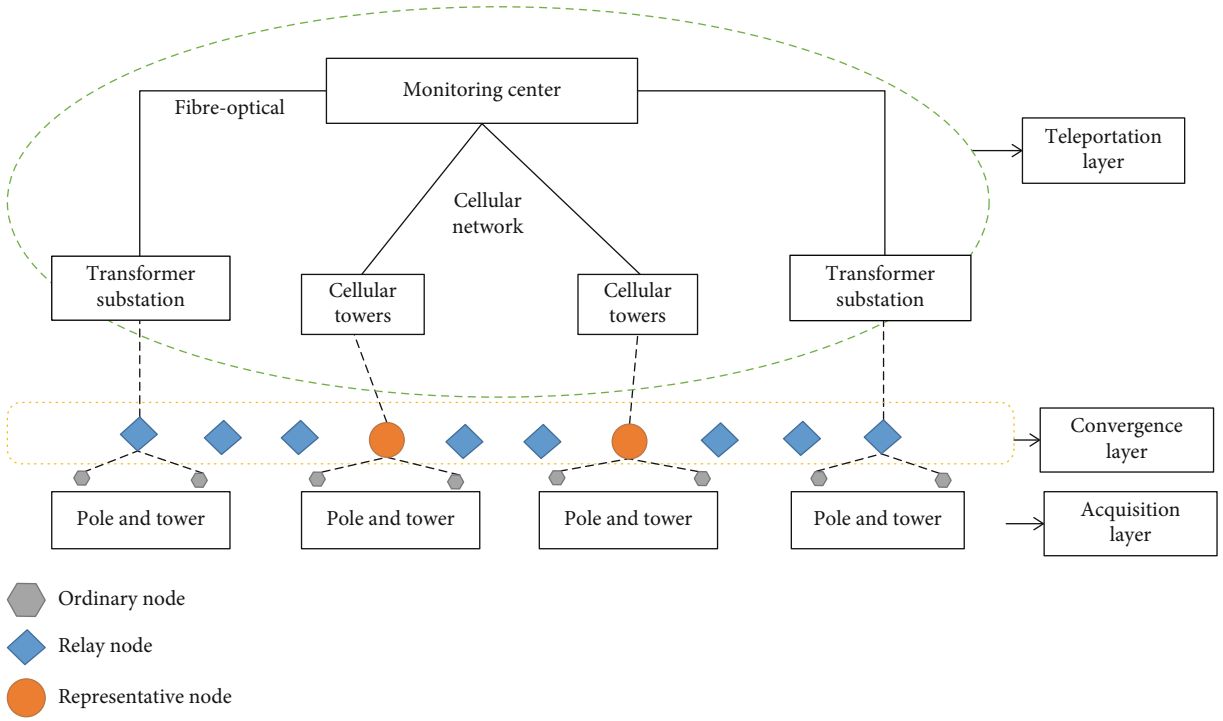


Figure 1: Hybrid hierarchical network structure.

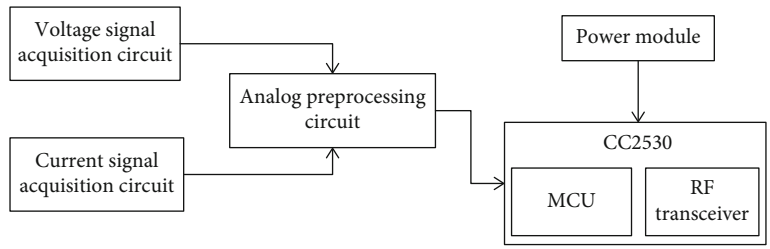


Figure 2: Hardware diagram of the acquisition layer.

in Equation (3) can be defined according to the protection setting value and the measured current.

$$I^{i}_{b\cdot\mathrm{dist}} = I^{i}_{b\cdot\mathrm{set}} - I^{i}_{b\cdot m}. \tag{3}$$

Here, $I^{i}_{b\cdot m}$ is the current measured by the backup protection on the i-side of the branch road L_b, $I^{i}_{b\cdot\mathrm{set}}$ is the setting value of the backup protection on the i-side of the branch road L_b, and $I^{i}_{b\cdot\mathrm{dist}}$ is the distance between $I^{i}_{b\cdot\mathrm{set}}$ and $I^{i}_{b\cdot m}$.

From Equation (3) and the action characteristics of the protection, it can be seen that when $I^{i}_{b\cdot\mathrm{dist}} > 0$, the$i$-side of the $L_b$will not have cascading trips, and when $I^{i}_{b\cdot\mathrm{dist}} < 0$, the i-side of the branch road L_b will have cascading trips, and when $I^{i}_{b\cdot\mathrm{dist}} = 0$, the i-side of the branch road L_b is at the boundary of cascading trips.

Similarly, if the j-side of branch road L_b is equipped with current-type backup protection, the cascading trips of branch road L_b will be caused after the protection action. The form of a similar equation (3) can be used to judge it, and the superscript i in Equation (3) can be replaced by j. For branch road L_b, either the i-side or j-side shows cascading trips; then, the branch will have cascading trips.

According to the judgment equation of branch road cascading trips given in Equation (3), all branch roads except the initial fault branch road in the power grid are considered, and then, the judgment of cascading trips at the power grid stratification plane can be further given. In fact, from the power grid stratification plane, according to the performance of the power grid cascading trips, the cascading trip of at least one branch road can be regarded as the cascading trips of the power grid. At this time, the state of the power grid can be called the state of the power grid cascading trips. If at least one branch road of the power grid is at the boundary of the cascading trips and the remaining branch roads are in the safe state except for those branch roads in the boundary state of the cascading trips, the state of the power grid can be called the critical state of the power grid cascading trips. Further, in addition to the state in which cascading trips occur in the power grid and the critical state, the state in which the power grid is in which cascading trips do not occur is the safe state.

To give a further mathematical description of the above state, the various types of backup protection on the remaining branch roads of the power grid except the initial faulted branch roads can be considered uniformly, and the variable C is used to represent I in Equation (3) uniformly, and all the backup protections are numbered uniformly. For any branch road L_b, the specific numbering can be done according to the following rules: if the i-side of the branch road L_b is equipped with current-type backup protection, it will be recorded as $C^{(1)}_{b\cdot\mathrm{dist}}$, and if the j-side of the branch road L_b is also equipped with current-type backup protection, it will be recorded as $C^{(2)}_{b\cdot\mathrm{dist}}$. After such treatment, the diagonal matrix of the following form can be given:

$$\mathbf{C} = \mathrm{diag}\left(C_{1\cdot\mathrm{dist}},\cdots,C^{(1)}_{b\cdot\mathrm{dist}}, C^{(2)}_{b\cdot\mathrm{dist}},\cdots,C^{(2)}_{m\cdot\mathrm{dist}}\right), \tag{4}$$

where m is the total number of branches in the power grid except for the initial fault branch roads. Equation (4) is written according to the current-type backup protection on both sides of each branch. If a branch road is only equipped with backup protection on one side, the corresponding elements can be removed from Equation (4).

After Equation (4) is given, according to the above analysis, when at least one branch road in the power grid meets $C^{(k)}_{b\cdot\mathrm{dist}} < 0$, the power grid is in the state of cascading trips. Here, k can be taken as 1 or 2 according to the situation. When Equation (5) is satisfied, the power grid is in a safe state.

$$C^{(k)}_{b\cdot\mathrm{dist}} > 0, \quad k = 1, 2,\ b = 1, 2, \cdots, m. \tag{5}$$

Further, when Equation (6) is satisfied, the power grid is just in the critical state of cascading trips.

$$\begin{cases} |\mathbf{C}| = 0, \\ C^{(k)}_{b\cdot\mathrm{dist}} > 0, \quad k = 1, 2,\ b = 1, 2, \cdots, m, \end{cases} \tag{6}$$

where $|\mathbf{C}|$ is the determinant value of matrix $\mathbf{C}$.

For the critical state described by Equation (6), this paper gives the following proof: let the set of all operating states of the power grid be represented by T, and let T_1 be the set of operations without any branch road cascading trips, T_2 be the set of the power grid in the critical state of cascading trips, and T_3 be the set where cascading trips occur but does not belong to the critical state. According to this division, it is easy to know that $T = T_1 \cup T_2 \cup T_3$, and $T_1 \cap T_2 = \emptyset$, $T_1 \cap T_3 = \emptyset$, and $T_2 \cap T_3 = \emptyset$. Let $\Omega 1$ be an operation state satisfying Equation (6); if $\Omega 1 \in T_1$, then all branch roads in the power grid must satisfy Equation (5), and $|\mathbf{C}|$ is not zero, which is contradictory to Equation (6), so there must be $\Omega 1 \in T_1$; if $\Omega 1 \in T_3$, then at least one branch road L_b satisfies $C^{(k)}_{b\cdot\mathrm{dist}} < 0$ according to the definition of T_3, which is also contradictory to Equation (6), so there must be $\Omega 1 \in T_2$.

4.2. Mathematical Model of the Security Level of the Power Grid for Cascading Trips. Among the various states of the power grid for cascading trips, the critical state is a key state. Obviously, in the actual operation, to ensure the safety level of the power grid for cascading trips, the power grid should be as far away from the critical state as possible. Through the analysis of the power flow equation of the power grid, it can be seen that the redistribution of power flow is mainly determined by the nodal injection power before the initial fault. Therefore, to keep the power grid away from the critical state, it is necessary to keep the nodal injection power away from the nodal injection power corresponding to the critical state.

Among all such nodal injection power combinations, the one closest to the current operating state of the power grid is a special combination. If the nodal injection power combination in the current operating state of the power grid is far away from this combination, then the nodal injection power in the current operating state of the power grid must be farther away from the remaining nodal injection power combinations that make the grid in the critical state of cascading trips.

If the combination of nodal injection power in the current operating state of the grid is represented by the vector $\mathbf{S}'$ and the combination of nodal injection power that makes the grid in the critical state of the cascading trips is represented by the vector $\mathbf{S}$, then the distance between the two can be represented by the following equation:

$$d(\mathbf{S}) = \left\|\mathbf{S}' - \mathbf{S}\right\|. \tag{7}$$

Here, $\|\mathbf{S}' - \mathbf{S}\|$ denotes the parametric number for $\mathbf{S}' - \mathbf{S}$.

Based on the above analysis, the vector $\mathbf{S}$ obtained by taking the minimum value of $d(\mathbf{S})$ is used as the combination of the nodal injection power that, makes the power grid in the critical state of cascading trips, and for the convenience of analysis, the minimum value of $d(\mathbf{S})$ is here denoted as $f(\mathbf{S})$, as shown in the following equation:

$$f(\mathbf{S}) = \min d(\mathbf{S}) = \min \left\|\mathbf{S}'\text{-}\mathbf{S}\right\|, \tag{8}$$

Here, $f(\mathbf{S}) > 0$, The higher the $f(\mathbf{S})$, the safer the power grid, so $f(\mathbf{S})$ reflects the level of security of the power grid for cascading trips; $f(\mathbf{S})$ can be represented as a security index. The mentioned method is security level prediction for cascading trips based on nodal injection power, which later in this paper will be called the SLP method.

From the process before and after the occurrence of cascading trips in the power grid, when the power injected into the power grid $\mathbf{S}$ meets Equation (8), it also needs to meet certain physical constraints, including various equation constraints and inequality constraints. The equation constraint condition shall include the power flow constraint relationship of the power grid before and after the initial failure, which can be abbreviated to the form shown in equation (9):

$$\begin{cases} h^0(\mathbf{x}) = 0, \\ h^a(\mathbf{x}) = 0. \end{cases} \tag{9}$$

Here, $\mathbf{x}$ is the state variable of the system. h^0 is the mapped relationship of the power flow constraint satisfied by the grid before the initial failure. h^a is the mapped relationship of the power flow constraint satisfied by the grid after the initial failure branch road L_a is removed.

For the relevant inequality constraints, these include generator output constraints, node voltage constraints, and line power constraints before and after the initial failure. They will be written in an abbreviated form as

$$m^0(\mathbf{x}) \le 0. \tag{10}$$

Summing up the previous analysis, you get the equation shown in the following equation:

$$\begin{cases} f(\mathbf{S}) = \min \left\| \mathbf{S}' - \mathbf{S} \right\|, \\ h^0(\mathbf{x}) = 0, \\ h^a(\mathbf{x}) = 0, \\ m^0(\mathbf{x}) \le 0, \\ |\mathbf{C}| = 0, \\ C_{b.\mathrm{dist}}^{(k)} \ge 0, \quad k = 1, 2b = 1, 2, \cdots, m. \end{cases} \tag{11}$$

The model provided in Equation (11) is an optimization model that describes the problem of how to find the closest state to the current operating state among the running states that can trigger cascading trips of the power grid, with the state represented by the nodal injection power. Obviously, the nodal injection power corresponding to the closest critical state to the current operating state of the power grid can be obtained using Equation (11), as well as the power grid's current operating state for cascading trip security level index.

4.3. Evaluation Idea for Power Grid Security Level. The existing various optimization algorithms have been widely applying in several fields, such as brain storm optimization algorithm (BSO) [14], memetic algorithm (MA) [15, 16], firefly algorithm (FA) [17, 18], and particle swarm optimization algorithm (PSO) [19]. Considering that the PSO is easy to implement programmatically, straightforward, and quick and suitable for solving complex optimization problems that are difficult to be solved by various classical optimization algorithms, this paper draws on the evolution particle swarm optimization (EPSO) of literature [20] to solve the model, whose iterative formula is shown below.

$$\begin{cases} \mathbf{v}_i^{k+1} = w\mathbf{v}_i^k + b1r1\left(\mathbf{P}_{\mathrm{best.i}} - \mathbf{y}_i^k\right) + b2r2\left(\mathbf{g}_{\mathrm{best}} - \mathbf{y}_i^k\right), \\ \mathbf{y}_i^{k+1} = \mathbf{y}_i^k + \mathbf{v}_i^{k+1}. \end{cases} \tag{12}$$

Here, $\mathbf{y}_i^k$ is the iterative position of particle i at the kth time; $\mathbf{v}_i^k$ is the iterative velocity of particle i at the kth time; $\mathbf{P}_{\mathrm{best}\cdot i}$ is the individual optimal position of particle i; $\mathbf{g}_{\mathrm{best}}$ is the population optimal position; w is the inertia coefficient, which decreases linearly from 0.9 to 0.1; b_1 and b_2 are acceleration constants, which are all set to 2; and r_1 and r_2 are [0,1] randomly distributed random numbers.

To cooperate with EPSO, all the equation constraints in Equation (12) are combined and written in the form $e(\mathbf{x}) = 0$, and all the inequality constraints in Equation (12) are combined and written in the form $k(\mathbf{x}) \ge 0$. Then, the following objective function with the penalty factor may be further given [21]. The constrained problem is transformed into an unconstrained problem by adding a penalty function. In this way, the solution is faster, and the operation is simple.

$$f'(\mathbf{S}) = f(\mathbf{S}) + \sum_i \frac{1}{\alpha_i} \left[\min\left(0, -e_i(\mathbf{x})\right)\right]^2 + \sum_j \frac{1}{\beta_j} \left[\min\left(0, -k_j(\mathbf{x})\right)\right]^2. \tag{13}$$

Here, $e_i(\mathbf{x})$ is the component i in $e(\mathbf{x})$, and $k_j(\mathbf{x})$ is the component j in $k(\mathbf{x})$. α and β are penalty factors, and the value of the penalty factor depends on the actual situation.

In this way, Equation (13) can be used to obtain the nodal injection power in the critical state according to the EPSO, as well as the corresponding grid security level index, where the particle position $\mathbf{y}$ corresponds to the nodal injection power vector $\mathbf{S}$.t$\mathbf{x}$in Equation (13) can be understood as an intermediate variable that satisfies the system constraint

To sum up the above, this paper mainly follows the following ideas to assess the security level of the power grid for the cascading trips, and its flow of ideas is shown in Figure 3.

Further, as seen in the previous analysis that after the security index $f(\mathbf{S})$ is calculated, a circle can be obtained with the current state of the power grid as the center of the sphere and the radius of the security index $f(\mathbf{S})$. When the operating state of the power grid is inside this sphere, the power grid is secure. For example, in a 2-node system, the nodal injection power of the grid is assumed to be active power, as shown in Figure 4.

In Figure 4, $\mathbf{P}_a$ is the nodal injection power combination of the current state of the grid, denoted by •; the nodal injection power combination of the critical state is denoted by $*$, and $\mathbf{P}_b$ is the nodal injection power combination of the critical state closest to $\mathbf{P}_a$; with $\mathbf{P}_a$ as the center of the circle and the distance between $\mathbf{P}_a$ and $\mathbf{P}_b$ as the radius to build circle D, the region within circle D is C_1, and the remaining region is C_2.

As shown in Figure 4, when the current operating state of the power grid is in zone C_1 but deviates from $\mathbf{P}_a$ if the power grid is hit by the initial failure and cascading trips have not occurred, you can use WSN to make further verification. The specific idea is as follows: when the actual power grid is working in zone C_1 but deviates from $\mathbf{P}_a$ and the initial failure does occur, the use of WSN quickly collects the current of each branch of the power grid and compares it with the protection of the fixed value, to determine whether the grid has cascading trips; if there are no cascading trips, then the method of this paper is reliable to determine whether cascading trips occur in the power grid. If not, it indicates that the method in this paper is reliable..

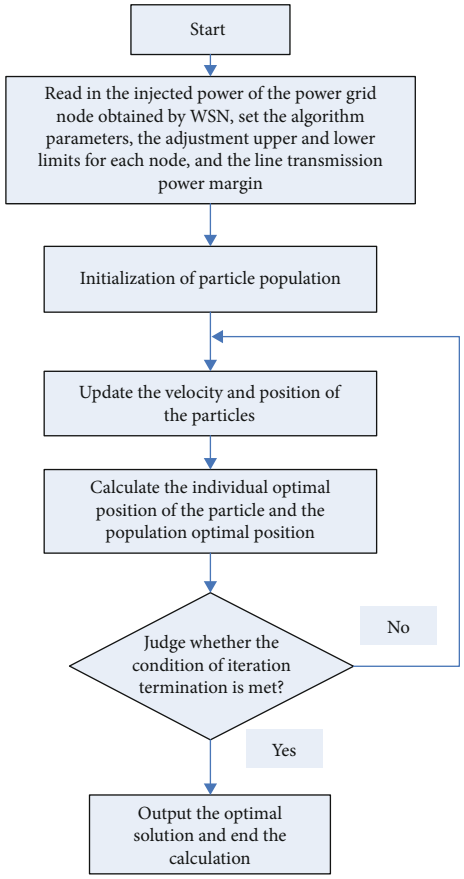


Figure 3: Algorithm flow.

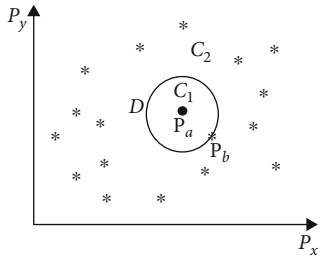


Figure 4: Security zone for the current operating state of the power grid.

5. Example Analysis

For illustration and validation of the previously proposed analytical model, this paper uses the IEEE39 node system for the algorithm demonstration; the wiring of the IEEE39 node system is shown in Figure 5. In the following example analysis, the results calculated in this paper are expressed as per-unit value, the reference capacity is considered as 100 MVA, and the reference voltage is consistent with the reference voltage of the IEEE36 bus data given in the literature [21].

According to the previous ideas in this paper, grid security assessment using WSN focuses on the analysis of the grid based on getting the nodal injection power, due to the former

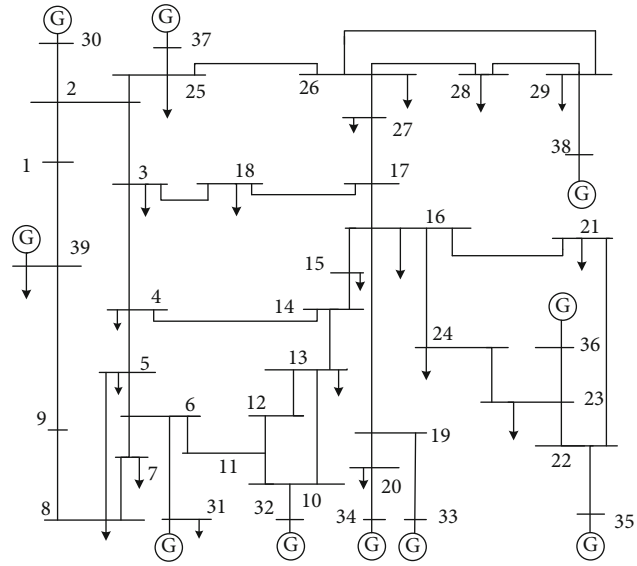


Figure 5: Diagram of IEEE39 node system.

advantages of WSN, which can obtain the nodal injection power of the entire power grid with a short delay and with high efficiency and accuracy. To focus on the verification of the evaluation method, this example assumes that the IEEE39 node system has configured WSN according to the structure shown in Figure 1, and the voltage phasor and current phasor collected through WSN are shown in Table 1 and Table 2, and their values are expressed as per-unit value. According to the obtained voltage and current, the nodal injection power is calculated according to Equation (2). Due to space limitations, the nodal injection power data will not be listed here. In the calculation, considering the actual characteristics of power flow, this paper mainly combines the PV node's active power, PQ node's active power, and reactive power to form the nodal injection power combination vector **S**.

With the solution idea of EPSO, it performed under the analysis program compiled in the Matlab environment. Suppose that the initial failure branch road is the branch road between node 5 and node 8, namely, branch road $L_{5,8}$. Each branch road of the system shown in Figure 5 configures the current-type backup protection, and the constant value of protection is 5.77 KA.

For the values of α_i and β_j in Equation (13), when they correspond to the constraint relationship in Equation (10), their values are taken as 0.01; when they correspond to the constraint relationship in Equation (6), α_i is taken as 0.0085, and β_j is taken as 0.0065. In the latter case, the smaller value of α_i and β_j is mainly because the constraint conditions corresponding to these two penalty factors are more closely related to the cascading trips. In the calculation, it is taken into account that $f(\mathbf{S})$ in Equation (13) is usually small and $|\mathbf{C}|$ is relatively large. For the convenience of analysis, the value of $f(\mathbf{S})$ is multiplied by $10e5$ in the procedure in this paper, while the value of $|\mathbf{C}|$ is divided by $10e5$, and the results below and the results of this paper will be given according to this requirement.

Table 1: Voltage phasor collected by WSN.

Node number	Voltage phasor	Node number	Voltage phasor	Node number	Voltage phasor
1	1.047	14	1.011	27	1.037
2	1.048	15	1.015	28	1.050
3	1.030	16	1.032	29	1.049
4	1.004	17	1.034	30	1.047
5	1.005	18	1.031	31	0.982
6	1.007	19	1.050	32	0.983
7	0.997	20	0.991	33	0.997
8	0.996	21	1.011	34	1.012
9	1.028	29	1.049	35	1.049
10	1.017	30	1.047	36	1.063
11	1.012	31	0.982	37	1.027
12	1	32	0.983	38	1.026
13	1.014	33	0.997	39	1.030

Table 2: Current phasor collected by WSN.

Branch head node number	Branch end node number	Current phasor	Branch head node number	Branch end node number	Current phasor
1	2	1.152	16	24	0.899
1	39	1.281	17	18	1.789
2	3	3.667	17	27	0.202
2	25	2.457	21	22	5.968
3	4	1.388	22	23	0.644
3	18	0.411	23	24	3.425
4	5	1.642	25	26	0.753
4	14	2.640	26	27	2.669
5	6	4.775	26	28	1.349
5	8	3.211	26	29	1.840
6	7	4.311	28	29	3.382
6	11	3.478	11	12	0.274
7	8	1.895	13	12	0.322
8	9	0.885	31	6	5.780
9	39	0.164	32	10	6.709
10	11	3.478	33	19	7.129
10	13	2.955	34	20	5.128
13	14	2.977	35	22	6.122
14	15	0.373	36	23	5.311
15	16	3.202	37	25	5.272
16	17	2.031	30	2	2.365
16	19	4.368	38	29	8.041
16	21	3.196	20	19	4.373

Figure 6 shows the comparison of the minimum f calculated by EPSO and PSO according to the same parameter setting. The total number of iterations is 200, and the horizontal ordinate in the figure indicates the number of iterations, while the vertical ordinate is the minimum $f'(\mathbf{S})$ obtained for each iteration.

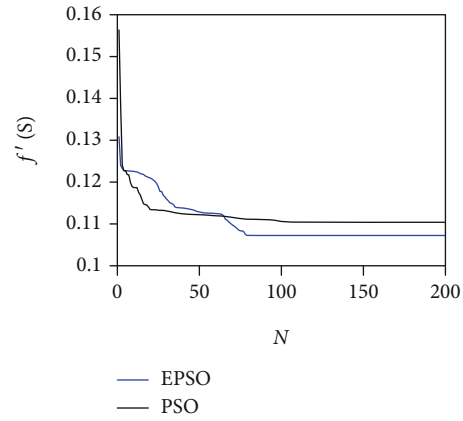


Figure 6: Calculation results of $f'(\mathbf{S})$ under different algorithms.

As can be seen in Figure 6, the EPSO results stabilize at the 86th calculation; i.e., the nodal injection power combination where the power grid is in the critical state of cascading trips is found. The minimum $f'(\mathbf{S})$ value is 0.5446, and the corresponding minimum $f(\mathbf{S})$ value is 0.4799. While the PSO results stabilize by the 66th calculation, the minimum $f'(\mathbf{S})$ is 0.1104 and the corresponding $f(\mathbf{S})$ is 0.0883. The comparison results show that EPSO has some advantages in calculating the minimum $f'(\mathbf{S})$, so EPSO is more suitable to be chosen. EPSO is used to obtain the nodal injection power in the $\mathbf{S}$ state corresponding to the minimum $f'(\mathbf{S})$, as shown in Table 3.

Table 3 shows the node injected power for the IEEE39 node system listed by the PQ node and PV node, respectively. For the PQ node, Table 1 lists the active and reactive power at its nodes, with active power on the left and reactive power on the right. For the PV node, Table 3 lists only the active power injected at the nodes. In Table 3, the unit of active power is MW, and the unit of reactive power is MVAR. If the power data in the table is positive, it means that the actual flow direction of this power is the outflow from the node. If it is negative, it means that the actual flow direction of this power is injected into the power grid from the node.

To verify the results obtained by EPSO in Figure 6, this paper uses the current and voltage collected by WSN to calculate the nodal injection power data according to Equation (2). Based on this data, by randomly modifying the power on each node, several nodal injection power states for comparison are obtained. Then, the $f'(\mathbf{S})$ of the power grid in these states are calculated, respectively, and they are compared with the minimum $f'(\mathbf{S})$ obtained by EPSO in Figure 6. The corresponding calculation results are shown in Figure 7. The ordinate of Figure 7 is the $f'(\mathbf{S})$ in the form of per-unit value, and the abscissa N_s is the serial number of the nodal injection power state, where the first injection power state is the corresponding one in Table 3.

In this paper, the injection power state of each node used for comparison in Figure 7 is formed as follows: according to the injection power state corresponding to the minimum f'

Table 3: The nodal injection power corresponding to the S state obtained by EPSO.

	Node number	P (MW)	Q (MVAR)
Load nodes (PQ)	3	-322	-2.405
	4	-500	-184
	7	-233.8	-84.001
	8	-522	-176
	12	-8.512	-88.002
	15	-330	-153
	16	-329	-32.293
	18	-158	-29.992
	20	-680	-103
	21	-274	-115
	23	-247.5	-84.601
	24	-308.6	92.198
	25	-224	-47.197
	26	-139	-16.995
	27	-281	-75.499
	28	-206	-27.591
	29	-283.5	-26.891
Power nodes (PV)	30	250	
	32	650	
	33	632	
	34	508	
	35	650	
	36	560	
	37	540	
	38	830	
	39	1000	

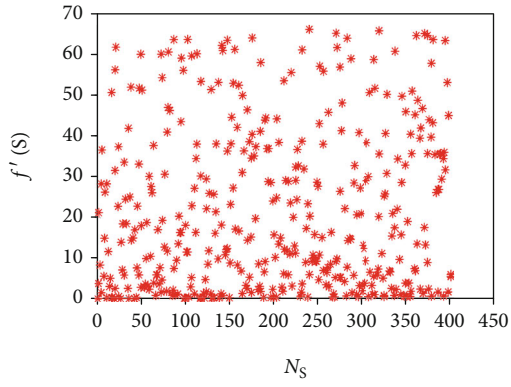


Figure 7: The $f'(\mathbf{S})$ corresponding to the injection power state of each node during verification.

$(\mathbf{S})$ in Figure 6, the power of each node i was modified according to the following equation:

$$S_i' = S_i + 0.1 \times q_1 \times q_2. \quad (14)$$

Here, q_1 is a random number uniformly distributed over the (0,1) interval, while q_2 is a number obtained by q_1. If $q_1 > 0.5$, then q_2 is taken as 1, and if $q_1 \le 0.5$, then q_2 is taken as -1.

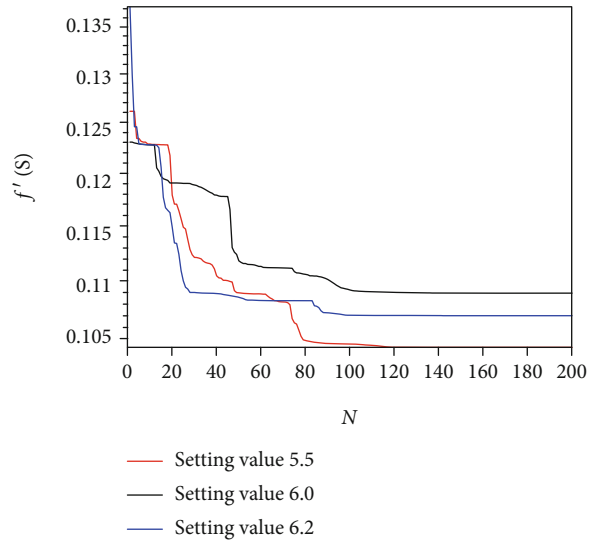


Figure 8: Calculation results of $f'(\mathbf{S})$ with different protection setting values.

After comparison, it can be seen from Figure 7 that the minimum $f'(\mathbf{S})$ obtained by EPSO in Figure 6 is also minimum in Figure 7, thus verifying the results in Table 3. The results obtained in this paper are similar to those in Figure 7 for several validations along the above lines, which are not drawn here due to the limitation of space. In conclusion, it can be seen from these comparisons that the minimum $f'(\mathbf{S})$ obtained by EPSO in Figure 6 is acceptable.

To further verify the method in this paper, EPSO is used to analyze a large number of examples for different relay setting values. The calculation of $f'(\mathbf{S})$ is similar to that in Figure 6. For the convenience of illustration, Figure 8 shows the calculation of three different protection settings. The meaning of ordinate and abscissa in Figure 8 is the same as that in Figure 6, and the protection setting values corresponding to each curve are shown in Figure 8.

As seen in Figure 8, the trend of $f'(\mathbf{S})$ changes during the iteration process is similar to that of Figure 6, and its value finally converges after a series of irregular decreases from the beginning of the iteration. In this paper, in addition to quantitative analysis for different protection setting values, for the initial fault, in addition to branch $L_{5,8}$ as the initial fault branch for calculation, for other branch roads composed of lines or transformers in the power grid, different branch roads are selected as the initial fault, and for different protection setting values for quantitative analysis and calculation, the calculation results are similar to Figure 8, while after a comparative analysis similar to Figures 6 and 7, the results are satisfactory, which shows that the model and solution ideas given in this paper are reasonable.

In the example, after calculating $f(\mathbf{S})$ and $f'(\mathbf{S})$, a safe region C_1 similar to the one shown in Figure 4 is obtained. For a given initial failure, assuming several operational states

are in region C_1, the current of each branch of the power grid is calculated by using the power flow calculation. According to Equation (5), it is determined that the power grid will not be cascading trips. Since the WSN can quickly and effectively collect the voltage and current of the power grid, the collected voltage and current are consistent with the calculation results of power flow. Therefore, it shows that WSN can be used to verify the previous calculation.

In short, through the above example, we can find that the corresponding $f'(\mathbf{S})$ and $\mathbf{S}$ can be obtained by using the method in this paper. The $f'(\mathbf{S})$ can help the operators to observe the distance between the current power grid and the critical state of the cascading trips, and $\mathbf{S}$ can alert the operators to the specific location of the closest critical state, thus providing a basis for the operator to avoid the power grid entering the critical position.

6. Conclusion

This paper studies the nodal injection power data of the power system based on WSN calculations and transmits these data to the monitoring center, and the monitors use the nodal injection power data to evaluate the security of the current power grid, and the paper gives the relevant analysis models and evaluation methods, and the main conclusions are as follows:

(1) As an ideal choice to deal with the new challenges of power grid parameter monitoring technology, WSN provides a more flexible and perfect solution for power grid parameter monitoring and can realize the centralized management of multitype power grid parameter monitoring. Using WSN, the power injection data of power grid nodes can be given quickly, accurately, and comprehensively. With the corresponding security level analysis algorithm, it can quickly and accurately evaluate the online security of the power system

(2) Since the cascading trips of the power grid are closely related to the action behavior of the relay protection, the mathematical expression of the cascading trips of the power grid can be given by the distillation of the action equation of the protection

(3) The SLP method mainly uses the nodal injection power of the grid to obtain the current operating state of the grid and the security evaluation after the migration of the operating state, which is simple and practical

(4) This paper presents a model to analyze the security level of the power grid for cascading trips, which can be used not only to calculate the security level of the power grid for cascading trips but also to calculate the initial critical operation state closest to the current operation state of the power grid, which provides a strong basis for avoiding the power grid entering the critical state and thus avoiding cascading trips

In a word, the method proposed in this paper can provide a reference for further research on cascading fault of the power grid and provide theoretical and technical support for the actual operation of the power grid.

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 is supported by the Scientific Research Development Foundation of the Fujian University of Technology under the grant GY-Z17149 and the Scientific and Technological Research Project of Fuzhou under the grant GY-Z18058.

References

[1] M. Zhou, Q. Liang, H. Wu, W. Meng, and K. Xu, "Wireless sensor networks for smart communications," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 4727385, 2 pages, 2018.

[2] D. Zhi-gang and S. Teng-fei, "Development status and application prospect of electronic current transformer," *Instrumentation Technology*, vol. 2019, no. 5, pp. 37–44, 2019.

[3] B.-S. Kim, H. S. Park, K. H. Kim, D. Godfrey, and K.-I. Kim, "A survey on real-time communications in wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 1864847, 14 pages, 2017.

[4] V. C. Gungor, B. Lu, and G. P. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 10, pp. 3557–3564, 2010.

[5] F. Fan, Q. Ji, G. Wu, M. Wang, X. Ye, and Q. Mei, "Dynamic barrier coverage in a wireless sensor network for smart grids," *Sensors*, vol. 19, no. 1, pp. 3557–3564, 2019.

[6] Z.-j. Bao, W.-j. Yan, and G. Wu, "Control of cascading failures in coupled map lattices based on adaptive predictive pinning control," *Journal of Zhejiang University-Science C(Computers & Electronics)*, vol. 12, no. 10, article 1258, pp. 828–835, 2011.

[7] J. Xu, X. Bai, and B. Huang, "Research on wide area cooperative precontrol system for cascading trips," *Power System Technology*, vol. 37, no. 1, pp. 131–136, 2013.

[8] M. Tian, X. Wang, Z. Dong et al., "Cascading failures of interdependent modular scale-free networks with different coupling preferences," *Epl*, vol. 111, no. 1, article 18007, 2015.

[9] Y. F. Wang, K. L. Gao, T. Zhao, and J. Qiu, "Assessing the harmfulness of cascading trips across space in electric cyber-physical system based on improved attack graph," *Proceedings of the CSEE*, vol. 36, no. 6, pp. 1490–1499, 2016.

[10] A. Hilmani, A. Maizate, and L. Hassouni, "Automated real-time intelligent traffic control system for smart cities using wireless sensor networks," *Wireless Communications and*

Mobile Computing, vol. 2020, Article ID 8841893, 28 pages, 2020.

[11] B. Wang, J. Sen, and Z. Shaomin, "Research on WSN topology and protocol for transmission lines monitoring," *Information Systems and Signal Processing Journal*, vol. 4, no. 1, 2019.

[12] B. Fateh, M. Govindarasu, and V. Ajjarapu, "Wireless network design for transmission line monitoring in smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1076–1086, 2013.

[13] L. I. Miao, C. Xiaobo, J. Xinchun et al., "Monitoring for overhead transmission lines based on WSN in smart grid," *Shaanxi Electric Power*, vol. 44, no. 10, pp. 1–5, 2016.

[14] X. Xue and J. Lu, "A compact brain storm algorithm for matching ontologies," *IEEE Access*, vol. 8, no. 8, pp. 43898–43907, 2020.

[15] X. Xue and Y. Wang, "Optimizing ontology alignments through a memetic algorithm using both MatchFmeasure and unanimous improvement ratio," *Artificial Intelligence*, vol. 223, no. 223, pp. 65–81, 2015.

[16] 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, 2016.

[17] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, pp. 2056–2115, 2020.

[18] X. Xue, "A compact firefly algorithm for matching biomedical ontologies," *Knowledge and Information Systems*, vol. 62, article 1443, no. 7, pp. 2855–2871, 2020.

[19] Y. Zhiqiang, H. Zhijian, and J. Chuanwen, "Economic dispatch and optimal power flow based on chaotic optimization," *Proceedings. International Conference on Power System Technology*, vol. 2002, no. 4, pp. 2313–2317, 2002.

[20] G. Shanghong and P. Tinglong, "Photovoltaic electricity generation power prediction based on similar day and LS-SVM with EPSO," *Transducer and Microsystem Technologies*, vol. 38, no. 3, pp. 40–46, 2019.

[21] S. Lian, S. Meng, and Y. Wang, "An objective penalty function-based method for inequality constrained minimization problem," *Mathematical Problems in Engineering*, vol. 2018, Article ID 7484256, 7 pages, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6639558, 6 pages
https://doi.org/10.1155/2021/6639558

Research Article

Design and Implementation of the Optimization Algorithm in the Layout of Parking Lot Guidance

Zhendong Liu, Dongyan Li, Yurong Yang, Xi Chen, Xinrong Lv, and Xiaofeng Li

School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China

Correspondence should be addressed to Zhendong Liu; liuzd2000@126.com

Received 31 December 2020; Revised 11 February 2021; Accepted 27 March 2021; Published 12 April 2021

Academic Editor: Xingsi Xue

The information guidance system for parking spaces in large- and medium-sized parking lots is not efficient at present. It tends to be difficult to find an empty parking space in parking lots in big cities. One of the problems is the large amount of calculation in the traditional Dijkstra algorithm. In this paper, an improved Dijkstra algorithm is presented and optimized to find the best parking path with the purpose of looking for the nearest free parking space based on the layout model in parking lot parking guidance. The experiments show that the improved Dijkstra algorithm can find the optimal parking space and the optimal parking path and improve the parking efficiency.

1. Introduction

With the continuous expansion of the scale of cities, the problem of parking difficulties has become increasingly prominent and traffic accidents frequently occur. The main reasons for the problem of "difficult parking" are as follows: the current urban parking spaces are in short supply; more importantly, in the process of searching for parking spaces, people can get less valuable parking space information; and there is a lack in parking space information management platform to guide vehicle drivers to park reasonably [1]. Solutions that help people to find the nearest parking lot have been put forward by researchers with the application of GPS technology/BeiDou technology becomes mature. The research focuses on outdoor parking guidance, but the indoor navigation technology has not been developed as it should be [2]. In fact, the indoor navigation technology and the outdoor navigation technology are essentially the same and they all need three kinds of technical support, namely, indoor positioning technology, indoor map, and path planning technology [3]. The research of indoor navigation and positioning technology has been paid more attention, and its application has gradually become popular. In the future, the construction of smart cities will be inseparable from indoor navigation and positioning technology. The development of indoor navigation and positioning technology with high precision, low cost, and universality and the realization of indoor and outdoor seamless navigation and positioning have always been the hot and difficult research topics at home and abroad [4]. Now, the development of indoor navigation has become a new direction for major enterprises to break through. One of the main applications of indoor navigation is to guide cars to park in parking lots. When the car owners arrives at the destination parking lot, they need to rely on the guidance of the managers in the parking lot or drive blindly to find the parking space, which wastes time and energy and also brings new problems: parade parking, resulting in congestion in the parking lot. It may take a long time for car owners to find appropriate parking spaces. It will greatly affect the parking of car owners when they encounter traffic congestion in the parking lot. Therefore, we need a parking space information management platform, which will be better if it has the function of guiding parking. It is an urgent problem that the guidance of parking spaces is imperfect and needs to be solved [5].

To address this challenge, we established an abstract data model of parking guidance based on the layout model in parking lot parking guidance, optimized this data model by using Dijkstra's improved algorithm, and finally found an optimal parking path with the lowest cost. With this method,

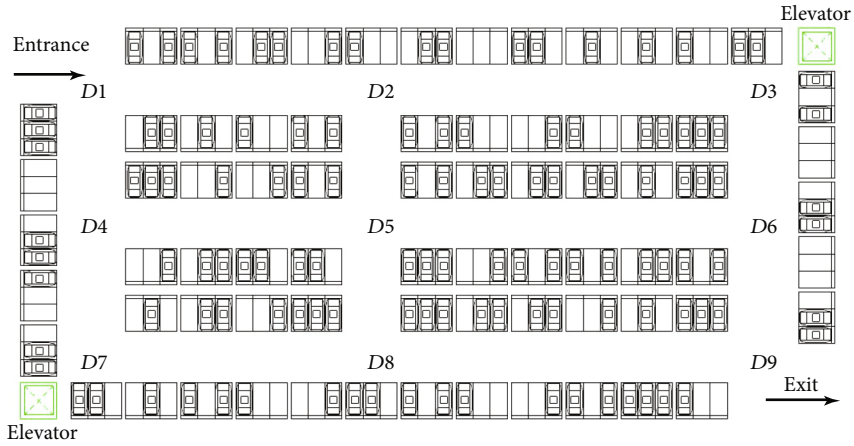


Figure 1: Parking space guidance layout diagram.

the car owners can find an idle parking space in the shortest time and parking efficiency and users' parking experience is greatly improved; the efficiency of the parking lot will also be improved.

2. System Structure and Parking Space Model Establishment

2.1. Parking Garage Information Guidance Management System. We designed and implemented an parking garage information guidance management system for intelligent parking. The main functions of the system includes parking space information collection, transmission, guidance control, and display statistics. Its main function is to monitor the occupancy situation of parking spaces in the parking lot and parking specifications in real time. There are many methods for monitoring the parking space status. For example, GPP and PGS2 systems adopt ultrasonic waves, while Siemens systems use parking sensors placed on the ground [6]. Monitoring vehicles by infrared sensors is one of the most sensible methods while considering that the system is mainly aimed at large- and medium-sized parking lots, and ultrasonic sensors are very sensitive to temperature changes and extreme air [7]. Therefore the system adopts infrared photoelectric switches. By arranging four infrared photoelectric switches in the parking space and adjusting the positions of infrared radio and television switches, the system can judge the state of vehicle parking irregularly if all four infrared photoelectric switches are covered by vehicles; it is standard. The working process of the parking garage information guidance management system is as follows: the infrared photoelectric switch on the parking space collects the specific information of the parking space, transmits the collected information, obtains the parking space status in the background, and selects all the empty parking spaces. When the owner enters the parking lot, there is a demand for finding the best empty parking space. The system finds the shortest path of the best empty parking space for the vehicle according to the guidance algorithm and publishes the information to the vehicle owner and guides him to park his vehicle as soon as possible. At the same time, the system will also detect the parking specification after the owner has parked the vehicle, which is mainly realized by setting reasonable infrared photoelectric switches. If the parking is standard, the system will feedback to the owner that the parking is successful. If the parking is not standardized, the system will prompt the owner of the vehicle to park the vehicle in a standardized way through the mobile phone. Intelligent parking solutions and intelligent systems provide a way to obtain parking lot information [8] and guide car owners to park faster and better and improve parking efficiency.

2.2. Model of the Parking Space Guidance Layout. Figure 1 is a typical parking space guidance layout diagram. The parking lot has only one exit and one entrance, and they are set up separately without affecting each other. The two pedestrian elevators are located in the middle of the entrance and exit, which are symmetrical as a whole. We discuss the optimization of the optimal parking path in this paper, taking this typical parking layout as an example. It can be regarded as a weighted graph; the cars to be parked, the intersections of lanes, and all free parking spaces are regarded as nodes; the paths in each driving direction are regarded as an edge; and the length of the driving road can be regarded as the weight of the edge. According to the path weights in the parking space guidance layout, the path optimization algorithm is used to calculate the weights of parking spaces and the best parking space and the corresponding parking path are selected to improve parking efficiency.

Many factors need to be considered in the selection of an empty parking space. The problems of the parking space itself are as follows: too close to the exit or entrance may cause safety problems and the empty parking space without cars parked next to it is better than the empty parking space with cars parked next to it [9]. There is also the issue of the owner's angle: the distance between the empty parking space and the pedestrian elevator and the distance between the empty parking space and the exit or entrance [10]. However, considering that the parking garage information guidance management system also involves guiding the car to the parking lot in the early stage and the owner has been driving the car for a period of time and the owner may be tired after

//P is the set of free parking spaces, S is the weight (path distance) of free parking spaces
1 calculate the distance between longitude and latitude of free parking spaces and longitude and latitude of the car M to be parked, the set of 5 free parking Spaces with the shortest distance is denoted as P and the free parking Spaces are marked as $P_i(I \in [0, 4])$.
2 Supposing that the shortest time shortest path set is T ($P \subsetneq T$), the corresponding weights (path distance) is $S(i)$, the set T is initialized to null and the weights $S(i)$ is initialized to 0, $i \in [0, 4]$;
3 While not all the elements in the set P enter the set T do
4 Select P_k, $P_k = \{P_i \in P - T \mid minS(i)\}$, P_k is the end point of the shortest path from the current car M, $T = T \cup \{P_k\}$;
5 Update the weight of the shortest path from car M to P_k, $S(k) = \min\{S(k), S(i) + S(i, k)\}$//$S(i, k)$ indicates the path distance between node i and node k.
6 End while
7 Output the final weights of all empty parking nodes in the empty parking space set P, and the berth Pi corresponding to $\min\{S(i)\}$ is the optimal berth.
8 The path corresponding to $W = \{M, \cdots, Dn, \cdots P_i\}$ is the optimal path from the car M to the optimal P_i.

Algorithm 1: ImDA(P, M).

arriving at the parking lot, so it is better to let the owner park the car in the shortest time. In this way, the parking efficiency of the parking lot is improved and the situation of car congestion is avoided. Set the free parking space in the parking lot as Pi, set the path distance from the parked vehicle to the free parking space as $S(i)$, and then, set the corresponding optimal parking path as $\min\{S(i)\}$.

3. Optimization Algorithm in the Layout of Parking Lot Guidance

In life, we are faced with many problems related to the shortest path. For example, it can reduce the cost of transporting unit materials and save freight expenses if we chose a reasonable transportation route under the condition of the existing traffic network. Thus, the problem of path optimization is crucial to our life.

3.1. Classical Shortest Path Algorithm and Its Disadvantages. The shortest path problem is a classical algorithm problem in graph theory, which is aimed at finding the shortest path between two nodes in a graph (composed of nodes and paths). Shortest path algorithms which are commonly used include the Dijkstra algorithm, Floyd algorithm, A$*$ (A Star) algorithm, BFS algorithm, and Johnson algorithm [11]. Among them, the Dijkstra algorithm is used to solve the shortest distance between a certain source point and other end points [12]; simply put that it is a typical single-source shortest path algorithm, which is neither DFS search nor BFS search, DFS occupies less memory but is slower, and BFS occupies more memory but is faster. The Dijkstra algorithm avoids these two problems. The algorithm is simple and easy to implement and has a high practical value. The Floyd algorithm is used to find the distance between any two points, which is different from the Dijkstra algorithm. In short, the Floyd algorithm is a multisource shortest path algorithm, which adopts the method of dynamic programming. The Floyd algorithm is simple and easy to implement, but by comparing their time complexity, Dijkstra algorithm's time complexity is generally $O(n^2)$, while Floyd algorithm's time complexity is generally $O(n^3)$. Therefore, the Dijkstra algorithm is faster than the Floyd algorithm, that is, the Dijkstra algorithm has more advantages in finding the shortest path of a single source. As we all know, finding free parking spaces in parking lots is a typical example of finding the shortest path of a single source. In addition, the Dijkstra algorithm is used to determine the shortest path and configure other conditions such as parking lane intersections, parking spaces, and their occupancy rates. The free parking space resources can be used more efficiently [13]. Therefore, this paper mainly improves the Dijkstra algorithm.

In the parking space guidance layout diagram, it will have a large amount of calculation when using the Dijkstra algorithm to find the shortest parking path and the efficiency of finding the best parking space will be very low, resulting in the problem of low parking efficiency. Therefore, the algorithm needs to be improved.

3.2. Improved Dijkstra Algorithm. There will be a sea of nodes and path when the traditional Dijkstra algorithm is applied to large parking lots, which makes the weighted graph more complicated. When the traditional Dijkstra algorithm is used to find the best parking path, the calculation is heavy and the parking efficiency is low. It is the simplest way to choose the nearest free parking space to the vehicle itself when we choose the optimal free parking space. Based on this idea, we design an optimization algorithm based on the Dijkstra algorithm in this paper. The basic idea of this optimization algorithm is as follows: obtain the latitude and longitude of free parking spaces and vehicles to be parked and calculate the distance between free parking spaces and vehicles to be parked by using the obtained latitude and longitude information; Euclidean distance is used here because it can better represent the true distance between two nodes. We select the nearest five free parking spaces which can be appropriately increased when the parking lot is as large as nodes and add them to the weighted graph, thus greatly reducing the number of nodes in the weighted graph. Then, the weights of the five free parking spaces are calculated (the path distance is taken as the weight). Finally, the parking space with the minimum value is selected to determine the best parking path. To facilitate the representation of parking paths, the lane intersections $D1$–$D9$ are introduced. Here is the improved Dijkstra algorithm:

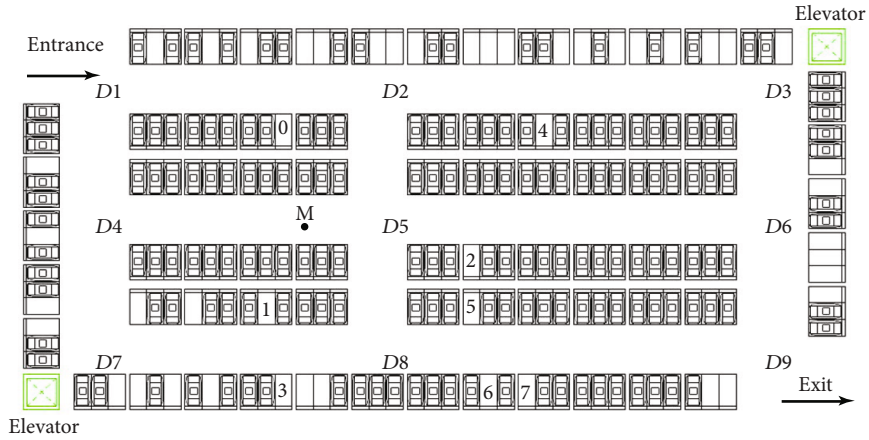


Figure 2: Parking space guidance layout at a certain time.

Table 1: Path distance in the layout.

Path	Path distance	Path	Path distance
$D1$–$D4$	7	P_0 – $D2$	6
$D4$ – $D7$	8	P_1 – $D8$	7
$D4$ – $D5$	15	P_2 – $D5$	5
$D5$ – $D6$	18	P_3 – $D8$	6
M – $D5$	5	P_4 – $D2$	8

Table 2: Parking space guidance weight and path.

Parking space	Path	Path distance (weight)
P_0	$M \longrightarrow D5 \longrightarrow D2 \longrightarrow P_0$	18
P_1	$M \longrightarrow D5 \longrightarrow D8 \longrightarrow P_1$	20
P_2	$M \longrightarrow D5 \longrightarrow P_2$	10
P_3	$M \longrightarrow D5 \longrightarrow D8 \longrightarrow P_3$	19
P_4	$M \longrightarrow D5 \longrightarrow D2 \longrightarrow P_4$	20

4. Experimental Analysis

Combined with the above improved Dijkstra algorithm, a specific example is selected for analysis. Figure 2 is a parking space guidance layout diagram of the parking lot at a certain time, which can clearly see the parking space occupation situation in the parking lot at this time. The following example takes Figure 2 as an example; assume that $P_0 \cdots \cdots P_7$ shown in the figure is the free parking space in the parking lot at this time (assuming that all but these seven parking spaces are occupied) and M represents the car to be parked.

According to the abovementioned optimization algorithm, firstly, the free parking spaces $P_0 \cdots \cdots P_7$ (represented by 0, 1, 2, 3, 4, 5, 6, and 7, respectively, in the figure) and the longitude and latitude of the car to be parked are obtained and the free parking spaces (the center points of rectangular parking spaces) are taken to obtain $C_0 \cdots \cdots C_7$; calculate the distance from M to $C_0 \cdots \cdots C_7$ and select the five points with the shortest straight line distance. The five points screened out in this example are P_0, P_1, P_2, P_3, P_4 (represented by 0, 1, 2, 3, and 4, respectively, in the figure), and these five points are the nodes that need to be added to the weighted graph.

Assuming that the measured distance (taking the width of a parking space as the unit distance) is as shown in Table 1 below, assume that each road is as wide as three parking widths, points $D1$–$D9$: two parking spaces from the far left of the road and one parking space from the right side of the road.

According to the abovementioned optimization algorithm for the parking space guidance layout diagram, the five points P_0, P_1, P_2, P_3, and P_4 closest to the car to be parked are screened out and the weights of these five points are calculated, namely, the path distance, which can also be said to be the driving distance of the route. The smaller the weight, the smaller the driving distance from the driving to the parking space, the simpler the parking, and the shorter the required time. The optimal parking path and weights from the car M to the free parking space are shown in Table 2.

It is not difficult to see that the parking guidance weight of parking space P_2 is the lowest, that is, the best parking space for car M is P_2 and the corresponding best parking path is $M \longrightarrow D5 \longrightarrow P_2$. After preliminary screening, the optional idle parking spaces are reduced and the number of nodes is greatly reduced. The shortest straight line distance between the source point and the end point of the car does not mean the shortest driving route. There may be the following special circumstances: the straight line distance is short, but it cannot be reached directly, so it takes a long distance to reach the parking space and the actual driving route is longer than other idle parking spaces. Therefore, parking spaces should be screened first, then, the shortest path of the screened parking spaces should be further calculated, and finally, the smallest weight should be selected as the best parking space. In this example, P_0 is the closest straight line distance from the source point (car M) to the free parking space; the corresponding parking path is $M \longrightarrow D5 \longrightarrow D2 \longrightarrow P_0$, and the weight (path distance) is 18. However, according to the above algorithm, the shortest parking path is $M \longrightarrow D5 \longrightarrow P_2$ and P_2 is the best parking space with a weight (path distance) of 10. By comparing the parking space P_0, obtained by the shortest straight line distance with the parking space P_2 obtained by the algorithm, it can be seen

that the optimized Dijkstra algorithm reduces the path distance, reduces the driving distance, and improves the parking efficiency, which further illustrates the superiority of the algorithm. A parking garage information guidance management system can guide owners to park the car faster and detect the parking conditions of vehicle owners through diffuse infrared photoelectric switches deployed in parking spaces. If it is detected that the car owners are parking irregularly, it can push the parking irregularity information to the user's mobile phone to remind the user to park the car regularly.

5. Conclusion

The traditional Dijkstra algorithm takes all empty parking spaces, parked cars, and lane intersections as nodes and adds all possible paths into the weighted graph, which is computationally intensive and inefficient. In view of this shortcoming of the traditional Dijkstra algorithm, we design an improved Dijkstra algorithm by limiting conditions, the number of nodes and the number of paths are reduced, the amount of computation is reduced, the computing efficiency is improved, and thus, the parking efficiency and parking experience are improved.

The main innovation of this paper is as follows: when looking for the best parking space, not considering the walking distance of the owner or the distance from the entrance and exit, then, the appropriate weight calculation method is selected to choose the parking space, which is different from other studies. We choose the car as the search center in this paper, which can ensure that the parking space is relatively close and optimal and ensure that the car is parked in the shortest time. In addition, it can also ensure that cars can search parking spaces anytime and anywhere in the parking lot, so as to find parking spaces that are close to themselves and convenient, with relatively few limitations.

However, there are many shortcomings in this paper that need to be further improved:

(1) It is necessary to strictly prove the improved Dijkstra algorithm to ensure that it is effective for graphs of large order, which is mainly aimed at the case of large parking lots

(2) Finding an unoccupied parking slot by the interested vehicle owners with the least overhead becomes an NP-hard problem bounded by various constraints [14]. Therefore, it is necessary to continue to improve the Dijkstra algorithm and improve the time complexity of the algorithm

(3) The calculation method of weights is relatively simple, which needs to be further improved and strictly compared with other methods, such as the neural network-based predictive control approach [15, 16]

(4) The parking space library information guidance management system is mentioned above, but this paper mainly introduces the Dijkstra algorithm for finding the best parking space [17, 18]. The description of the system is relatively few, and the system still needs further improvement. For example, after finding the best parking space, you can rely on the display at the intersection to indicate the driving direction or that the roadside indicator lights up to directly guide the car to the parking space, so as to realize a more intelligent parking system

Data Availability

All data are available within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Our work was supported by the NSFC under Grant nos. 61672328 and 61672323, and the research is also supported by the Science and Research Plan of the Luoyang branch of Henan Tobacco Company, no. 2020410300270078.

References

[1] Y. Yang, *Design and Implementation of Intelligent Parking Guidance System*, Jilin University, 2018.

[2] S. Liao, K. Du, Y. Zhao et al., "Scheme design of parking navigation system in indoor parking lot," *Value Engineering*, vol. 38, no. 7, pp. 162–164, 2019.

[3] T. Lei and Y. Chu, "Analysis on the current development of indoor navigation technology," *Modern Business Trade Industry*, vol. 41, no. 24, p. 153, 2020.

[4] W. Gao, C. Hou, W. Xu, and X. Chen, "Research progress and prospect of indoor navigation and positioning technology," *Journal of Navigation and Positioning*, vol. 41, no. 24, p. 153, 2020.

[5] W. Wang and Z. Gai, "Design of parking guidance path in parking lot based on Dijkstra algorithm," *Network Security Technology & Application*, vol. 9, pp. 52-53, 2018.

[6] J. Hanzl, "Parking information guidance systems and smart technologies application used in urban areas and multi-storey car parks [J]," *Transportation Research Procedia*, vol. 44, pp. 361–368, 2020.

[7] M. Bachani, U. M. Qureshi, and F. K. Shaikh, "Performance analysis of proximity and light sensors for smart parking [J]," *Procedia Computer Science*, vol. 83, pp. 385–392, 2016.

[8] R. Singh, C. Dutta, N. Singhal, and T. Choudhury, "An improved vehicle parking mechanism to reduce parking space searching time using firefly algorithm and feed forward back propagation method [J]," *Procedia Computer Science*, vol. 167, pp. 952–961, 2020.

[9] W. Li, Q. Li, and S. Yu, "Parking guidance system of parking lot using internet of things technology," *Automation & Instrumentation*, vol. 4, pp. 234-235, 2015.

[10] Y. Zhang and S. Tian, "Application of Dijkstra optimization algorithm in parking space guidance system," *Computer Measurement and Control*, vol. 22, no. 1, pp. 191–193, 2014.

[11] S. Wang and Z. Wu, "Improved Dijkstra shortest path algorithm and its application," *Computer Science*, vol. 39, no. 5, pp. 223–228, 2012.

[12] W. Yan and W. Wu, *Data Structure (C Language Version)*, Tsinghua University Press, 2012.

[13] J. Fan, "Predicting vacant parking space availability: an SVR method with fruit fly optimisation," *IET Intelligent Transport Systems*, vol. 12, no. 10, pp. 1414–1420, 2018.

[14] S. Saharan, N. Kumar, and S. Bawa, "An efficient smart parking pricing system for smart city environment: a machine-learning based approach [J]," *Future Generation Computer Systems*, vol. 106, pp. 622–640, 2020.

[15] J.-H. Shin, J.-G. Kim, and H.-B. Jun, "Dynamic control of intelligent parking guidance using neural network predictive control," *Computers & Industrial Engineering*, vol. 120, pp. 15–30, 2018.

[16] H. Liu, Y. Wang, and N. Fan, "A hybrid deep grouping algorithm for large scale global optimization," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 6, pp. 1112–1124, 2020.

[17] S. Liu, Y. Wang, W. Tong, and S. Wei, "A fast and memory efficient MLCS algorithm by character merging for DNA sequences alignment," *Bioinformatics*, vol. 36, no. 4, pp. 1066–1073, 2020.

[18] S. Wei, Y. Wang, Y. Yang, and S. Liu, "A path recorder algorithm for multiple longest common subsequences (MLCS) problems," *Bioinformatics*, vol. 36, no. 10, pp. 3035–3042, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6627588, 12 pages
https://doi.org/10.1155/2021/6627588

Research Article

Attention Mechanism-Based CNN-LSTM Model for Wind Turbine Fault Prediction Using SSN Ontology Annotation

Yuan Xie,[1] **Jisheng Zhao**,[2] **Baohua Qiang**,[1] **Luzhong Mi**,[2] **Chenghua Tang**,[1] **and Longge Li**[1]

[1]*School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*Beijing Huadian Tianren Electric Power Control Technology Co., Ltd., Beijing 100039, China*

Correspondence should be addressed to Baohua Qiang; qiangbh@guet.edu.cn

Received 30 December 2020; Revised 22 February 2021; Accepted 10 March 2021; Published 27 March 2021

Academic Editor: Xingsi Xue

The traditional model for wind turbine fault prediction is not sensitive to the time sequence data and cannot mine the deep connection between the time series data, resulting in poor generalization ability of the model. To solve this problem, this paper proposes an attention mechanism-based CNN-LSTM model. The semantic sensor data annotated by SSN ontology is used as input data. Firstly, CNN extracts features to get high-level feature representation from input data. Then, the latent time sequence connection of features in different time periods is learned by LSTM. Finally, the output of LSTM is input into the attention mechanism module to obtain more fault-related target information, which improves the efficiency, accuracy, and generalization ability of the model. In addition, in the data preprocessing stage, the random forest algorithm analyzes the feature correlation degree of the data to get the features of high correlation degree with the wind turbine fault, which further improves the efficiency, accuracy, and generalization ability of the model. The model is validated on the icing fault dataset of No. 21 wind turbine and the yaw dataset of No. 4 wind turbine. The experimental results show that the proposed model has better efficiency, accuracy, and generalization ability than RNN, LSTM, and XGBoost.

1. Introduction

In recent years, with the development of human beings, the exploitation and utilization of petroleum and fossil fuels have promoted the development of various fields. However, due to the over exploitation of these resources, the nonrenewable energy is reduced gradually, which makes the energy crisis and climate change become an urgent problem [1]. Countries of the world begin to turn their attention to renewable energy. Wind energy as a renewable energy has the characteristics of clean and easy to use, which has aroused great concern in the world. At present, although wind energy develops rapidly, how to predict accurately the occurrence of wind turbine fault and reduce effectively the maintenance cost of wind farm are still urgent problems to be solved.

Sensor technology has been applied widely in various fields. A large number of sensors are also installed on the wind turbine equipment to collect the operation data of the wind turbine. Based on the collected mass operation data, wind turbine state data analysis method is used to predict the fault of wind turbine. This method analyzes the state data of the wind turbine from multiple layers and excavates the potential valuable fault information of data by intelligent algorithm, realizing the fault prediction of the wind turbine. In recent years, there have been some related research results. There are mainly fault prediction models based on traditional machine learning methods. Kusiak and Verma [2] used data mining technology to analyze the bearing overtemperature fault and established a bearing fault prediction model to predict the fault. Although the overtemperature fault can be predicted accurately, the false alarm rate is relatively high, which increases the burden of maintenance personnel. Kusiak and Li [3] layer upon layer processed SCADA data of wind turbine. Based on the processed data, the wind fault prediction model was constructed and predicted successfully for wind turbine fault, but the accuracy of the model needs to be improved. Zhong et al. [4] proposed a rapid data-driven fault diagnostic method, which integrates data preprocessing

and machine learning techniques. In this method, fault features are extracted by using the modified Hilbert-Huang transforms and correlation techniques. Pairwise-coupled sparse Bayesian extreme learning machine is applied to build a real-time multifault diagnostic system. The experimental results show that the method can diagnose quickly the fault, but the generalization ability of the model is poor. Chen and Zhang [5] proposed a fault prediction method of wind turbine blade cracking based on RF-LightGBM algorithm. In this method, firstly, random forest algorithm ranks the importance of features to select important features. Then, the classification model is trained by the data after feature selection and optimized by K-fold cross validation. Finally, the method predicted successfully the wind turbine blade fault. Wang et al. [6] put forward the XGBoost algorithm-based fault prediction model of wind turbine main bearing. Wu et al. [7] presented a fault diagnosis method of wind turbine based on ReliefF and XGBoost algorithm. Hsu et al. [8] proposed a novel fault diagnosis technique for wind turbine gearbox. While statistical process control was applied to fault diagnosis, random forest and decision tree algorithms were employed to construct the predictive models for wind turbine anomalies. Fan and Tang [9] proposed a wind turbine pitch anomaly recognition system based on AdaBoost-SAMME. The proposed models in literature [5–9] have high accuracy, but the generalization ability of these models needs to be improved. Wang et al. [10] proposed a fault diagnosis method of wind turbine gearbox based on Riemannian manifold, which is fast in model training, but the accuracy of the model is not high. Zhang et al. [11] proposed a wind turbine fault diagnosis method based on the Gaussian process metamodel, which has excellent performance. But the model has high dependence on dataset, which leads to poor generalization ability of the model.

With the development of deep learning, there are many models for wind turbine fault prediction based on deep learning. Chen et al. [12] proposed a fault diagnosis method of wind turbine gearbox based on wavelet neural network. This method can predict faults accurately. But it cannot learn deep features in data. Lu et al. [13] proposed a wind turbine planetary gearbox fault diagnosis method based on self-powered wireless sensor and deep learning. This method has high accuracy, but it cannot learn the temporal relationship between data. Kavaz and Barutcu [14] proposed a wind turbine sensor fault detection method based on artificial neural network. This method combines with SCADA system for data acquisition, which makes the model have the advantages of low cost, but the model lacks the ability to process time series data. Chen et al. [15] proposed a model of the relationship between root cause and symptom of wind turbine fault based on Bayesian network to predict wind turbine fault. However, the complexity of the model increases exponentially with the increase of parent nodes of Venn diagram, which makes the training time too long. Shi et al. [16] proposed an intelligent fault diagnosis method of wind turbine bearing based on convolution neural network. This method can extract features effectively, but the accuracy of the method needs to be improved. Zhang et al. [17] proposed a fault diagnosis method of wind turbine gearbox based on 1DCNN-PSO-SVM. This method has a good performance in feature extraction, but it lacks the ability to learn time series features. In order to solve the problems of low efficiency and poor accuracy of manual detection method for wind turbine blade defect, Zhang and Wen [18] proposed an improved Mask R-CNN detection method for wind turbine blade defect. Firstly, ResNet-50 combined with FPN is used to generate feature map. Then, it is input into RPN to screen out the ROI. Finally, the size of the feature map is determined by ROIAlign, which is input into the full connection layer network for blade defect detection. The experimental results show that this method has fast detection efficiency and high accuracy. However, this method only considers the characteristics of a single data graph and does not mine the temporal relationship between data. Chang et al. [19] proposed a concurrent convolution neural network for fault diagnosis. The method has high accuracy and strong generalization ability. However, the network structure parameters and calculational speed of this method need to be further optimized. Jiang et al. [20] proposed a fault diagnosis model of wind turbine gearbox based on multiscale convolution neural network. The proposed MSCNN incorporates multiscale learning into the traditional CNN architecture. It improves greatly the feature learning ability and diagnosis performance of the model. But the model cannot learn the long-term correlation between the data. Yin et al. [21] proposed a temperature fault early warning method for wind turbine main bearing based on Bi-RNN, but there are problems of gradient disappearance and gradient explosion for processing time series data. Yin et al. [22] proposed a fault diagnosis method of wind turbine gearbox based on the optimized LSTM neural network with cosine loss. However, this method cannot learn features directly from the original vibration sign. Zheng et al. [23] put forward the fault prediction method for wind turbine gearbox based on K-means clustering and LSTM. Although it can process effectively time series data, it inputs directly the original features of the wind turbine into the model for training, resulting in model training too long. On this basis, a fault prediction method for wind turbine based on deep trust network was proposed, which has a good effect on feature extraction of wind turbine data, but it has the problem of big error [24]. Lei et al. [25] presented an end-to-end LSTM model for fault prediction. The model uses the data fusion strategy of IDENTITY function to extract multisensor features. LSTM captures long-term dependencies through recurrent behaviour and gates mechanism. The experimental results show that this method is able to do fault classification effectively from raw time series signals collected by single or multiple sensors. However, the strategy used in the data fusion stage of this method needs to be improved.

In accordance with shortcomings of traditional machine learning in processing time series data and advantages of CNN [26] in feature extraction and LSTM [27] in processing time series data, an attention mechanism-based CNN-LSTM model for wind turbine fault prediction is constructed. The contributions of this method are summarized as follows:

(i) This method proposes a joint training mode of CNN and LSTM, which can process effectively the time series data and ensure the training speed. In addition,

considering that the irrelevant features of data will lead to the degradation of the model performance, therefore, attention mechanism [28] is used to redistribute feature weights to ensure the performance of the model and improve the generalization ability of the model

(ii) In order to further eliminate the influence of irrelevant features and improve the construction speed of the model, random forest algorithm [29] is used to select features in the data preprocessing stage, which further improves the generalization ability of the model

The rest of this paper is organized as follows. In Section 2, the data acquisition process of wind turbine is introduced. Section 3 introduces the model of this paper. Section 4 presents the experimental procedure and experimental results. Section 5 is the conclusion including summary of the method and future work.

2. Data Acquisition

In view of the isomerization of the data transmitted by the sensor in the form and description information, in this paper, SSN ontology annotation method is used to format and unify sensor data of wind turbine. The specific process of ontology annotation method is shown in Figure 1.

Firstly, the method needs to analyze the sensor data of wind turbine to find its concept and properties. Secondly, it is necessary to analyze the structure of SSN ontology and compare sensor data with SSN ontology structure to extract useful information. Based on these, the annotation information of sensor data can be obtained by the semiautomatic annotation mapping method. Then, the annotation information of data is generated into a mapping file in XML format by R2RML mode. The mapping file is used mainly to store the annotation information of wind turbine sensor data. Then, the mapping file of wind turbine sensor data is transformed into ontology instance to generate RDF semantic sensor data by the semantic conversion algorithm of sensor data. Finally, the RDF semantic sensor data is stored in HBase by Spark Streaming.

Based on the ontology semantic annotation method of SSN, SCADA system is used to collect data. The icing fault dataset and yaw fault dataset of a wind farm in Yunnan Province are collected. According to the SSN ontology structure diagram as shown in Figure 2, the corresponding wind turbine features are generated. Among them, the icing fault data involves 26 features, as shown in Table 1; the yaw fault data involves 28 features, as shown in Table 2.

3. Introduction to the Model

3.1. Model Structure. The structure of the attention mechanism-based CNN-LSTM wind turbine fault prediction model is shown in Figure 3. The model can be divided into three parts. In the first part, the processed data are input into CNN and the high-level feature representation is obtained by feature extraction of CNN. In the second part, the output of CNN is input into LSTM; the deeper temporal relationship between features is obtained by LSTM. In the third part, attention mechanism is introduced to acquire more related information of fault. This model is also called as the CLA model.

For a given training dataset $D = \{(x_i, y_i) \mid i = 1, \cdots, p\}$, x_i represents the input sample data, $x_i \in R^q$, where q is the number of features, and $y_i \in \{0, 1\}$ is the label of the i-th sample data. The sample data is input into CNN, and the high-level feature representation is obtained by extracting the original features:

$$J_i = f\left(\omega \times x_{i:i+g-1} + b\right), \tag{1}$$

where ω is the convolution kernel, g is the size of convolution kernel, $i : i + \mathrm{g} - 1$ is the i-th feature to the $(i + g - 1)$-th feature, and b is the offset term. After the calculation of convolution layer, the characteristic matrix J is acquired:

$$J = \left[c_1, c_2, \cdots, c_{n-g+1}\right]. \tag{2}$$

Then, the maximum pooling technique [30] is used to process the local characteristic matrix J of the fault to retain the key information of the features and reduce the parameters; finally, the local optimal solution is as follows:

$$M = \max\left(c_1, c_2, \cdots, c_{n-g+1}\right) = \max\{J\}. \tag{3}$$

Then, the M vector is connected to form the H vector by the full connection layer:

$$H = \{M_1, M_2, \cdots, M_n\}. \tag{4}$$

The output H of convolution network is taken as the input of LSTM. After receiving the output of CNN, the forgetting gate, memory gate, and output gate are calculated by the hidden state h_{t-1} of the last time and the current input x_t to form a memory unit, which runs through all processes. It can retain important information and remove unimportant information; the specific process is as follows:

(1) Firstly, the data passes through the forgetting gate. The sigmoid unit in the forgetting gate generates a vector between 0 and 1 by calculating h_{t-1} and x_t. According to this vector, LSTM can determine what information needs to be retained and what information needs to be discarded in memory unit C

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{5}$$

where f_t is the forgetting gate, σ is the activation function, and h_{t-1} and x_t are the inputs.

(2) The next step is to determine which part of the new information is added to the memory unit, which involves two operations. Firstly, input gate is obtained by calculating h_{t-1} and x_t; input gate can determine which part of the information needs to be updated. Then, let h_{t-1} and x_t pass through a

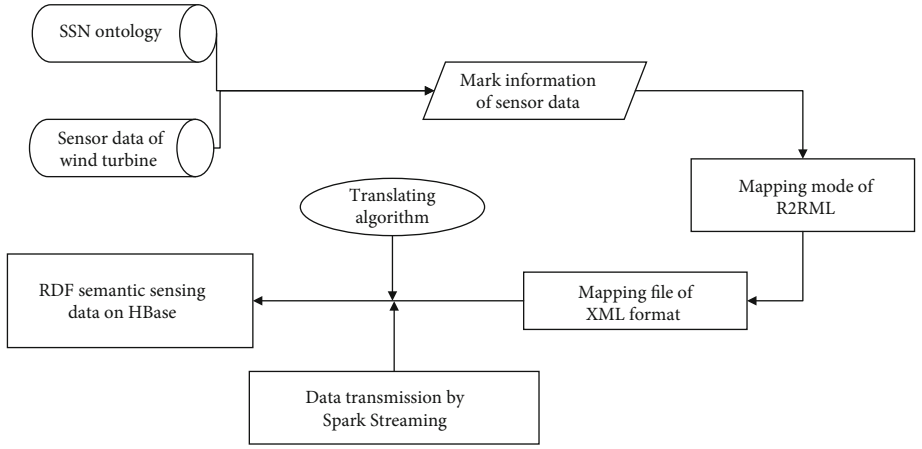


Figure 1: The process of SSN ontology annotation.

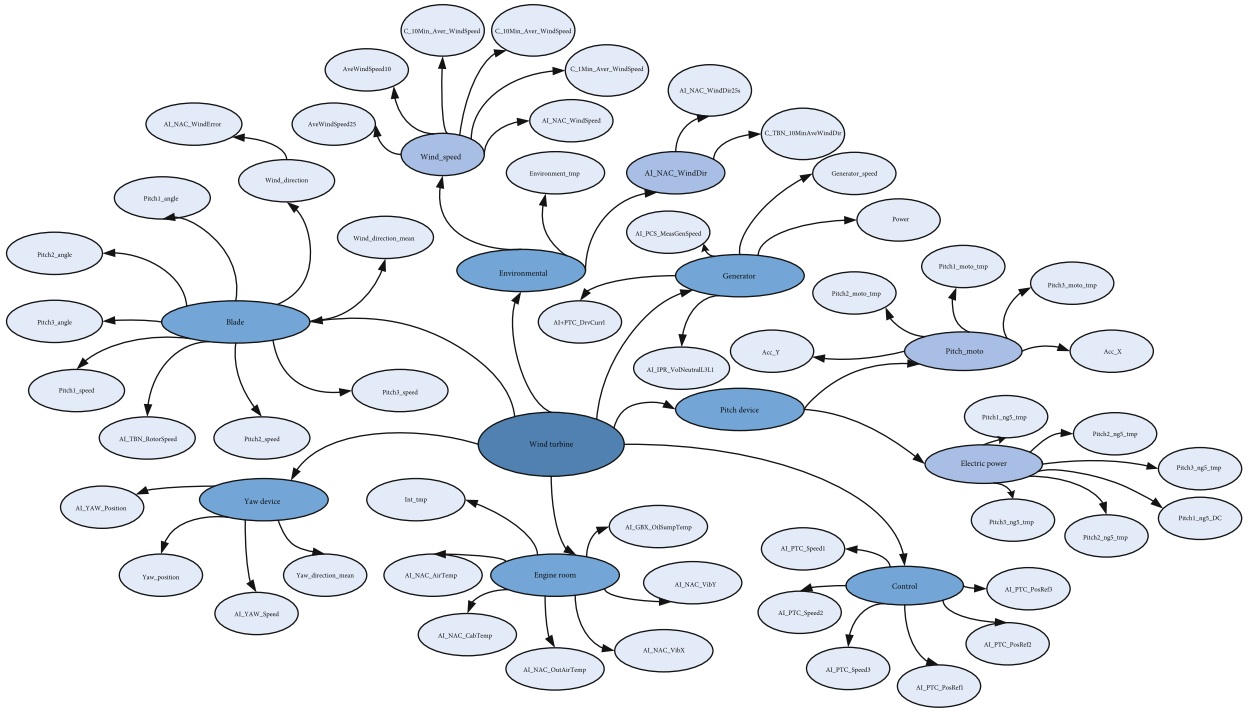


Figure 2: SSN ontology structure diagram.

Table 1: Original characteristics of wind turbine icing fault.

Features	Features	Features	Features
wind_speed	pitch1_angle	pitch2_moto_tmp	pitch2_ng5_tmp
generator_speed	pitch2_angle	pitch3_moto_tmp	pitch3_ng5_tmp
power	pitch3_angle	acc_X	pitch1_ng5_DC
wind_direction	pitch1_speed	acc_Y	pitch1_ng5_DC
wind_direction_mean	pitch2_speed	environment_tmp	pitch1_ng5_DC
yaw_position	pitch3_speed	int_tmp	
yaw_speed	pitch1_moto_tmp	pitch1_ng5_tmp	

Table 2: Original characteristics of wind turbine yaw fault.

Features	Features	Features	Features
aveWindSpeed25	AI_NAC_WindDir25s	AI_TBN_RotorSpeed	AI_PTC_PosRef1
aveWindSpeed10	C_TBN_10MinAveWindDir	AI_NAC_Position	AI_PTC_PosRef2
AI_NAC_WindSpeed	AI_GBX_OilSumpTemp	AI_YAW_Speed	AI_PTC_PosRef3
C_10Min_Aver_WindSpeed	AI_NAC_WindError	AI_PTC_Speed1	AI_NAC_VibX
C_15Min_Aver_WindSpeed	AI_NAC_AirTemp	AI_PTC_Speed2	AI_NAC_VibY
C_1Min_Aver_WindSpeed	AI_NAC_CabTemp	AI_PTC_Speed3	AI_PTC_DrvCurr1
AI_NAC_WindDir	AI_NAC_OutAirTemp	AI_PCS_MeasGenSpeed	AI_IPR_VoltNeutralL3L1

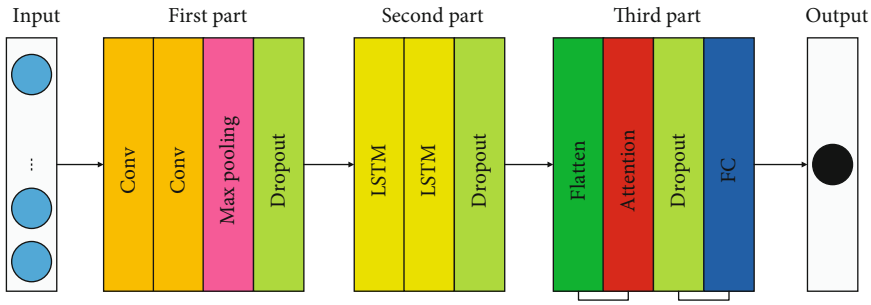


Figure 3: Structure of the CLA model.

tanh layer to get new candidate memory units $\tilde{C}_t$; this information of candidate memory may be updated into the memory unit

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= \tanh\,(W_C \cdot [h_{t-1}, x_t] + b_C), \end{aligned} \tag{6}$$

where i_t is the input gate, tanh is the activation function, and $\tilde{C}_t$ is the candidate memory unit.

(3) LSTM will update memory unit; memory unit C_{t-1} will be updated to memory unit C_t. The process of updating is as follows: firstly, forget some information of old memory unit through the forgetting gate. Secondly, increase part of the information of candidate memory unit $\tilde{C}_t$ through input gate. Finally, get a new memory unit C_t

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{7}$$

where C_t is the new memory unit.

(4) Finally, the LSTM determines which state characteristics need to be output through the output gate; let the input h_{t-1} and x_t pass through the sigmoid layer of the output gate to get a judgment condition, and then let the memory unit pass through the tanh layer to get a vector between -1 and 1, which is multiplied by the judgment condition of the output gate to get the final output of the memory unit

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{8}$$

$$h_t = o_t * \tanh\,(C_t), \tag{9}$$

where O_t is the output gate and h_t is the final output.

In the appeal of Equations (1) to (9), $W_{f'} W_{i'} W_{c'} W_o$ is the weight matrix and $b_{f'} b_{i'} b_{c'} b_o$ is the offset vector.

(5) The data is processed by LSTM to get the output $H_t = [h_1, \cdots, h_t]$ of each time step; then, input H_t into the attention module. Through the attention mechanism, different weights can be assigned to the fault characteristics of wind turbine. The formulas are as follows:

$$w_t = \tanh\,(H_t), \tag{10}$$

where h_t is the input and w_t is the target weight.

Then, the softmax function is used to probabilistically affect the attention weight:

$$P_t = \frac{\exp\,(w_t)}{\sum_{t=1}^{m} \exp(w_t)}, \tag{11}$$

where P_t is the weight probability vector.

The generated attention weight is assigned to the corresponding hidden layer state code h_t:

$$a_t = \sum_{t=1}^{m} P_t \cdot h_t, \tag{12}$$

where a_t is the weighted average of h_t and P_t is the weight.

Input: Status data of wind turbine after processed: $D = \{(x_i, y_i) \mid i = 1, \cdots, p\}$
Output: Fault prediction results of wind turbine.
- FC is the full connection layer;
- max is the maximum pool layer;
- σ is the sigmoid activation function;
- ω is the convolution kernel, g is the size of convolution kernel, $i : i + g - 1$ are the features from i to $i + g - 1$ and b is the offset term;
- W_o, W_f is the weight matrix, b_o is the offset vector, h_{t-1} is the hidden state of the previous time, x_t is the input of the current time, C_t is the memory unit of the current time;
- y is the prediction result of the model and $y^{(i)}$ is the real label of the sample.

Training:
Initialization: Initialize all parameters in the model;
For p in n do:
(1) Deep level feature is extracted based on CNN: $X_i = FC\{\max\{\sigma(\omega \times x_{i:i+g-1} + b)\}\}$;
(2) LSTM integrates the fault characteristics of each time segment into a unified sequential fault feature: $h_t = \sigma(W_o[h_{t-1}, x_t] + b_o) * \tanh(C_t)$
(3) The attention mechanism allocates the weights and the full connection layer generates the prediction results: $a_t = \sum\sigma(\tanh(H_t)) * h_t$, $y = FC(W_f \cdot a_t)$;
(4) Calculate the loss value of the model $L = \sum_{i=1}^{N} y^{(i)} \log y'^{(i)} + (1 - y^{(i)}) \log(1 - y'^{(i)})$, then adjust the model parameters.
End.

Algorithm 1: CLA model algorithm.

Finally, the results are output through the full connection layer:

$$y_t = \sigma(W_{\text{final}} \cdot a_t), \tag{13}$$

where y_t is the predication results, W_{final} is the weight matrix, and σ is the sigmoid activation function.

3.2. Model Training and Testing. Algorithm 1 shows pseudocode of CLA model algorithm. Firstly, all the trainable parameters in the model are initialized. Secondly, the training dataset $D = \{(x_i, y_i) \mid i = 1, \cdots, p\}$ is put into CNN through the input layer of the model. CNN uses convolution layer and pool layer to get the deep feature matrix $X = [X_1, \cdots, X_i]$ of data. Then, the output of CNN is used as the input of LSTM. LSTM can further learn the feature of wind turbine data at multiple continuous times to obtain the time series dependence between features. The output of each time step for LSTM recorded as $H_t = [h_i, \cdots, h_t]$ is input to the attention mechanism module to redistribute the feature weight. Finally, according to the weighted features a_t, the full connection layer generated the prediction results y. According to the error between the predicted results and the real values, the model updates the trainable parameters. The model judges whether the training times reach the upper limit of the maximum number of iterations. If not, the number of iterations is increased by 1 and the training process is repeated. Otherwise, the trained model can be obtained.

The training and testing flow of the CLA model is shown in Figure 4. Firstly, the processed data are divided into training dataset and test dataset. Secondly, the training dataset is input into the constructed model and the features of data are extracted by convolution neural network. Then, the results are input into LSTM for time series feature transformation; the fault characteristics of each time segment are integrated into a unified sequence fault features; besides,

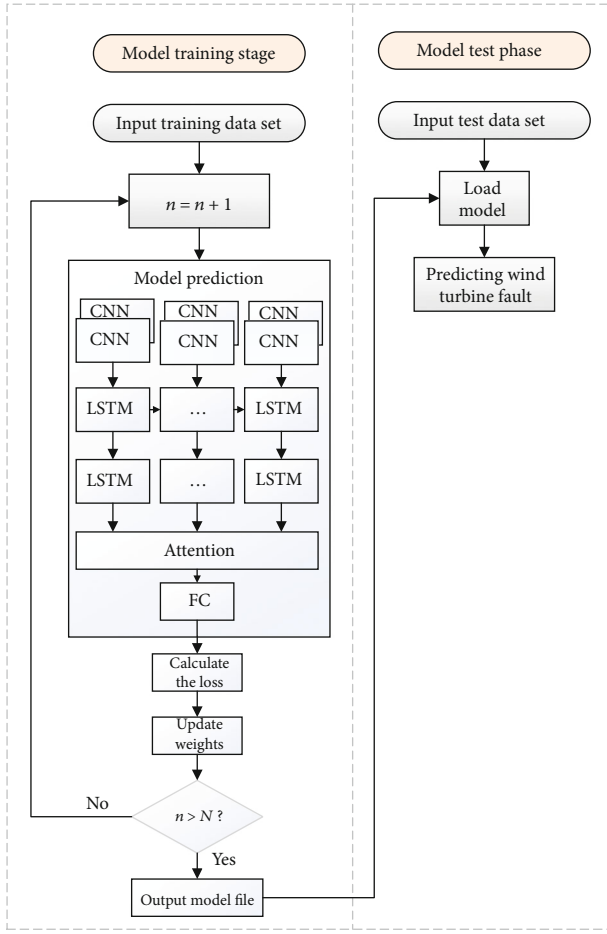


Figure 4: Process of the CLA model training and testing.

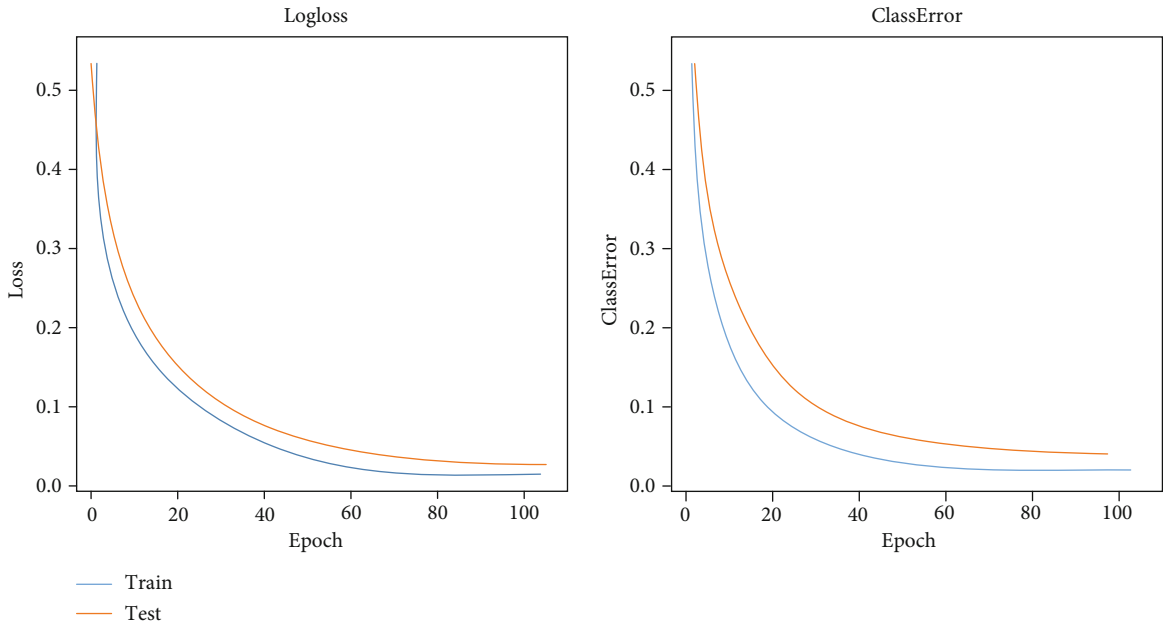


Figure 5: Logloss and ClassError of the CLA model on ice fault dataset of No. 15 wind turbine.

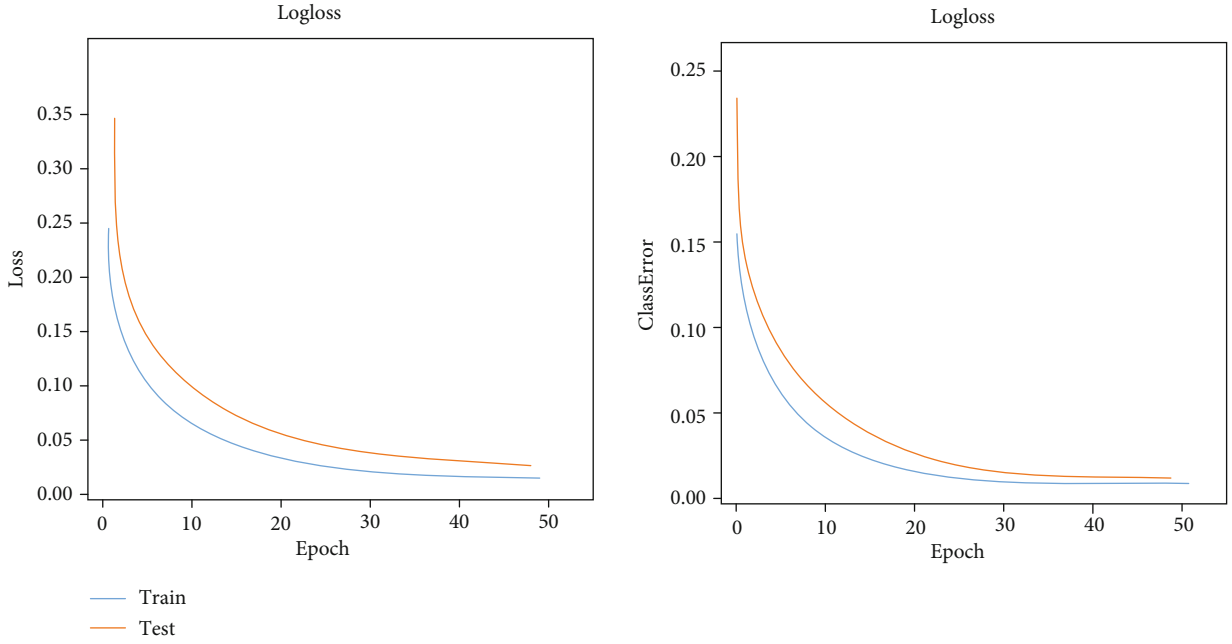


Figure 6: Logloss and ClassError of the CLA model on yaw fault dataset of No. 3 wind turbine.

attention mechanism is introduced in the model. Finally, the prediction results are output through the full connection layer and the model parameters are constantly updated according to the loss value. In this experiment, the number of iterations is used as the end condition of training; when the model reaches the preset number of iterations, the fault prediction model is obtained. In the test phase, firstly, the trained model is loaded. Then, the test dataset is input into the model for fault prediction. Finally, the experimental results are obtained.

The CLA model integrates CNN, LSTM, and attention mechanism. Among them, CNN can reduce model parameters by sharing convolution kernel parameters; thus, it can accelerate the calculation speed of the model. Through memory unit, LSTM can deal with the long time sequence dependence of data and solve the problem of gradient disappearance caused by too long time step. Attention mechanism can make the model pay more attention to the features that have a high correlation with the fault. Besides, attention mechanism also can reduce the influence of nonimportant

Table 3: Experimental results of the model on the test datasets of No. 15 and No. 3 wind turbine.

Algorithm	Wind turbine icing fault test dataset No. 15 wind turbine				Wind turbine yaw fault test dataset No. 3 wind turbine			
	A	*P*	*R*	*F*1	*A*	*P*	*R*	*F*1
CLA	0.9646	0.9730	0.9634	0.9712	0.9821	0.9805	0.9754	0.9819
LSTM	0.9388	0.9227	0.9580	0.9400	0.9793	0.9803	0.9781	0.9792
RNN	0.7986	0.8083	0.7833	0.7956	0.9672	0.9609	0.9738	0.9673
XGBoost	0.9805	0.9875	0.9733	0.9804	0.9855	0.9906	0.9852	0.9873

Table 4: Experimental results of the model on the dataset of No. 21 and No. 4 wind turbine.

Algorithm	Wind turbine icing fault dataset No. 21 wind turbine				Wind turbine yaw fault dataset No. 4 wind turbine			
	A	*P*	*R*	*F*1	*A*	*P*	*R*	*F*1
CLA	0.7492	0.8171	0.7421	0.7591	0.7709	0.8143	0.7429	0.7737
LSTM	0.7242	0.7279	0.7163	0.7220	0.7455	0.8000	0.6546	0.7201
RNN	0.6793	0.6793	0.6792	0.6793	0.7248	0.7637	0.6511	0.7029
XGBoost	0.7038	0.7120	0.6847	0.6981	0.7038	0.7120	0.6847	0.6981

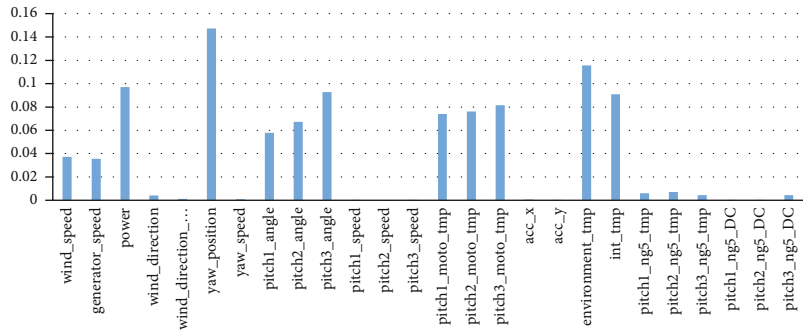


Figure 7: Importance of icing fault characteristic.

features, so as to improve the generalization ability and accuracy of the model.

4. Experiment and Analysis

The construction of the CLA model is based on Keras deep learning framework. In the experiment of wind turbine icing fault prediction, a two-layer convolution neural network and a two-layer LSTM neural network are used, where the number of neurons in each layer of convolution neural network and LSTM neural network is set to 64. In the experiment of wind turbine yaw fault prediction, a two-layer convolution neural network and a two-layer LSTM neural network are also used, where the number of neurons in each layer of convolution neural network is set to 32 and the number of neurons in each layer of LSTM neural network is set to 16. In order to prevent overfitting, dropout is introduced into the model and the value of dropout is set to 0.5.

4.1. Selection of Evaluation Indexes. In the CLA model, sigmoid is used as the activation function and binary_crossentropy is used as the loss function. We select the commonly used evaluation indicators in the field of fault prediction, including accuracy (*A*), precision (*P*), recall (*R*), and *F*1 value (*F*1).

Table 5: Characteristics of wind turbine icing fault after screening.

Features	Score
yaw_position	0.147249
environment_tmp	0.115538
power	0.097059
pitch3_angle	0.092650
int_tmp	0.090929
pitch3_moto_tmp	0.081600
pitch2_moto_tmp	0.076074
pitch1_moto_tmp	0.073881
pitch2_angle	0.067191
pitch1_angle	0.057764
wind_speed	0.037283
generator_speed	0.035642

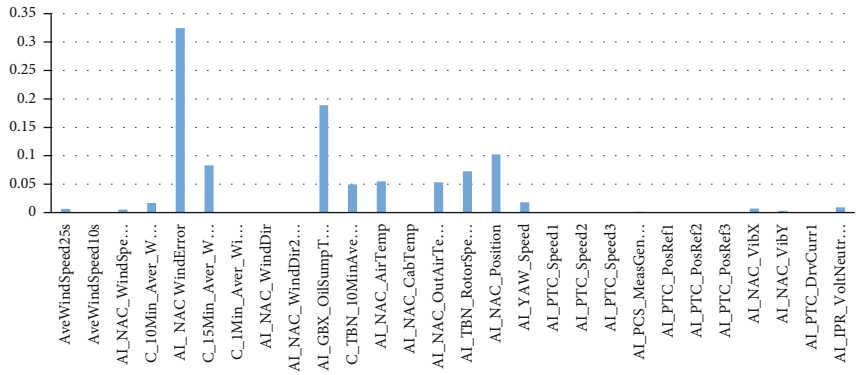


Figure 8: Importance of yaw fault characteristics.

Accuracy refers to the ratio between the number of samples classified correctly by the model and the total number of samples for a given test dataset:

$$A = \frac{TP + TN}{TP + FP + TN + FN}. \tag{14}$$

Precision refers to how many of the samples predicted as positive by the model are real positive samples:

$$P = \frac{TP}{TP + FP}. \tag{15}$$

Recall rate indicates how many positive samples in the dataset are predicted correctly:

$$R = \frac{TP}{TP + FN}. \tag{16}$$

$F1$ value represents the combined value of precision and recall rate. Because P and R are a pair of contradictory variables, it is difficult to simultaneously improve them in the experimental process. So, it is necessary to comprehensively evaluate them. Therefore, $F1$ value is proposed to reconcile them; $F1$ value is the comprehensive average of P and R. The closer the precision rate and the recall rate are, the greater the $F1$ is:

$$F1 = 2\frac{P \cdot R}{P + R}. \tag{17}$$

In Equations (14)–(17), TP is the number of samples that predict the wind turbine samples in normal state as normal state, FP is the number of samples that predict the wind turbine samples in fault state as normal state, FN is the number of samples that predict the wind turbine in normal state as failure state, and TN is the number of samples that predict the wind turbine in fault state as failure state.

4.2. Experimental Scheme. The hardware environment of the experiment is Intel Xeon e5-2698v4 (20 Core) processor and 128 G of memory, the operating system is 64 bit Ubuntu 16.04, Python 3.6 is used as software programming language, and PyCharm 2017.1.2 is used as a software development tool. The icing fault dataset of No. 21 and No. 15 wind turbine and the yaw fault dataset of No. 4 and No. 3 wind turbine are used in the experiment.

Table 6: Characteristics of wind turbine yaw fault after screening.

Features	Score
AI_NAC WindError	0.324536
AI_GBX_OilSumpTemp	0.188854
AI_NAC_Position	0.102554
C_15Min_Aver_WindSpeed	0.082901
AI_TBN_RotorSpeed	0.072518
AI_NAC_AirTemp	0.055106
AI_NAC_OutAirTemp	0.053102
C_TBN_10MinAveWindDir	0.049441
AI_YAW_Speed	0.018063
C_10Min_Aver_WindSpeed	0.016823

In order to verify the effectiveness of the proposed wind turbine fault prediction model in dealing with time series data and generalization problems, two groups of experiments are carried out, using RNN [21], LSTM [25], and XGBoost [6] algorithm as a comparison.

The first group of experiments, the data of No. 15 wind turbine and No. 3 wind turbine, is, respectively, divided into training dataset and corresponding test dataset. Two groups of training datasets are used directly to train the CLA model and other reference models; then, the corresponding test datasets are used to verify the model.

The second group of experiments, the two CLA models and other reference models trained in the first group, was tested on the data of No. 21 wind turbine and No. 4 wind turbine; the experimental results were compared and analyzed.

4.3. Experimental Results and Analysis. As shown in Figures 5 and 6, Logloss and ClassError of the CLA model on the dataset of Experiment 1 are shown, respectively. It can be found from the graph that Logloss and ClassError decrease continuously during the process of model iteration, which indicates that the model is converging with the increase of iteration times.

On the two datasets of Experiment 1, the results of the CLA model and other comparison algorithm models are shown in Table 3. On the test dataset of No. 15 wind turbine

Table 7: Experimental results after feature screening.

Algorithm	Wind turbine icing fault dataset No. 21 wind turbine				Wind turbine yaw fault dataset No. 4 wind turbine			
	A	*P*	*R*	*F*1	*A*	*P*	*R*	*F*1
CLA	0.7776	0.7899	0.7565	0.7728	0.8622	0.8228	0.9234	0.8702
LSTM	0.7664	0.7310	0.7431	0.7531	0.8401	0.8025	0.9023	0.8495
RNN	0.6855	0.6571	0.7756	0.7115	0.7751	0.7912	0.7475	0.7687
XGBoost	0.7345	0.7286	0.7571	0.7426	0.7345	0.7286	0.7571	0.7426

icing fault, the accuracy rate of the CLA model is 96.46%. Compared with the LSTM and RNN algorithm models, the accuracy rate of the CLA model is improved, respectively, by 2.58% and 16.60% and that of the other three indicators *P*, *R*, and *F*1 of the CLA model is also the best. Besides, the four indicators of the CLA model are all above 96%. But compared with the XGBoost algorithm model, the accuracy of the CLA model is reduced by 1.59%. On the test dataset of No. 3 wind turbine yaw fault, the accuracy of the CLA model is 98.21%. Compared with the LSTM and RNN algorithm models, the accuracy rate of the CLA model is improved, respectively, by 0.28% and 1.49% and that of the other three indicators *P*, *R*, and *F*1 of the CLA model is also the best. Besides, the four indicators are all above 96%. But the accuracy of the CLA model is reduced by 0.34% compared with the XGBoost algorithm model.

In Experiment 2, the trained model in Experiment 1 is applied, respectively, to the dataset of No. 21 wind turbine and No. 4 wind turbine; the experimental results are shown in Table 4. On the ice fault dataset of No. 21 wind turbine, the accuracy rate of the CLA model is 74.92%, which is 2.50%, 6.99%, and 4.54%, respectively, higher than the LSTM, RNN, and XGBoost algorithm models, and that of the other three indicators *P*, *R*, and *F*1 of the CLA model is also the best. On the yaw fault dataset of No. 4 wind turbine, the accuracy rate of the CLA model is 77.09%, which is 2.54%, 4.61%, and 6.71%, respectively, higher than the LSTM, RNN, and XGBoost algorithm models, and that of the other three indicators *P*, *R*, and *F*1 of the CLA model is also the best.

According to the experimental results of Experiment 1 and Experiment 2, the CLA model has the best performance in *A*, *P*, *R*, and *F*1 compared with the LSTM and RNN algorithm models; the results show that the temporal relationship of fault features can be learned to improve the accuracy and generalization ability of the model by fusion of CNN and LSTM and introducing attention mechanism. It proves the effectiveness of the CLA model. However, on the test dataset of No. 15 and No. 3 wind turbine, the accuracy rate of the CLA model is reduced, respectively, by 1.59% and 0.34% compared with the XGBoost algorithm model. Because the CLA model needs more data for deep learning compared with XGBoost, due to the lack of data, the model is over fitted, which makes the performance of the CLA model worse than the XGBoost algorithm model on the test dataset of No. 15 and No. 3 wind turbine, but the CLA model is more able to mine the deep relationship of time series data, so the generalization ability of the CLA model is better than XGBoost.

5. Optimization of CLA Model

In order to further improve the generalization ability of the CLA model, the original features of the icing fault dataset of No. 15 and No. 21 wind turbine are input into the random forest algorithm; the original features are screened through many experiments. Finally, the histogram of feature importance is obtained as shown in Figure 7. Based on this histogram, 12 features of the highest correlation degree with the icing fault of the wind turbines are selected, as shown in Table 5.

In the same way, random forest algorithm is used in the yaw fault dataset of No. 3 and No. 4 wind turbine; the feature importance histogram is obtained as shown in Figure 8. Based on this histogram, 10 features of the highest correlation degree with the yaw fault of the wind turbines are selected, as shown in Table 6.

Two feature datasets of No. 15 and No. 3 wind turbine screened by random forest algorithm were input into the CLA model for training; the trained model is used to test the feature dataset of No. 21 and No. 4 wind turbine screened by random forest algorithm for verifying the screening effect. The experimental results are shown in Table 7. For the icing fault dataset of No. 21 wind turbine after screening, the CLA model is still the best and the accuracy rate is improved by 2.84%. For the yaw fault dataset of No. 4 wind turbine after screening, the CLA model is still the best and the accuracy rate is improved by 9.13%. It shows that feature screening by random forest algorithm can improve effectively the generalization ability of the model.

6. Conclusions

In this paper, an attention mechanism-based CNN-LSTM model for wind turbine fault prediction is proposed. The model is trained by the icing fault dataset and yaw fault dataset of wind turbine annotated by SSN ontology. The trained model can predict accurately the occurrence of wind turbine fault. The experimental results show that the model is more effective and better than some of the current mainstream models. In this method, the model can learn effectively the deep-seated temporal characteristics among samples and focus on the features of high correlation with wind turbine fault through the joint training of CNN-LSTM and the introduction of attention mechanism. These strategies improve successfully the accuracy and generalization ability of the model. In addition, the generalization ability of the model can be further improved by using the random forest algorithm in the data preprocessing stage. In the later stage,

we will continue to study how to improve the model so that the prediction accuracy and generalization ability of the model can be further improved. Besides, we will also research how to introduce more excellent machine learning algorithms into the model for adapting to the complex working environment of wind turbine; it means that the model will be applied truly to practical engineering.

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 work was partially supported by the National Natural Science Foundation of China (61762025 and 62062028), the Guangxi Key Research and Development Program (AB18126053, AB18126063, and AD18281002), the Natural Science Foundation of Guangxi of China (2017GXNSF AA198226, 2019GXNSFDA185007, 2019GXNSFDA185006, and 2018GXNSFAA294058), the Guangxi Key Science and Technology Planning Project (Nos. AA18118031 and AA18242028), the National Energy Technology Environmental Protection Group Co., Ltd (IKY.2019.0002), the Innovation Project of GUET Graduate Education (2019YCXS051 and 2020YCXS052), and the Guangxi Colleges and Universities Key Laboratory of Intelligent Processing of Computer Image and Graphics (No. GIIP201603).

References

[1] M. Bhattacharya, S. R. Paramati, I. Ozturk, and S. Bhattacharya, "The effect of renewable energy consumption on economic growth: evidence from top 38 countries," *Applied Energy*, vol. 162, pp. 733–741, 2016.

[2] A. Kusiak and A. Verma, "Analyzing bearing faults in wind turbines: a data-mining approach," *Renewable Energy*, vol. 48, pp. 110–116, 2012.

[3] A. Kusiak and W. Y. Li, "The prediction and diagnosis of wind turbine faults," *Renewable Energy*, vol. 36, no. 1, pp. 16–23, 2011.

[4] J. H. Zhong, J. Zhang, J. J. Y. Liang, and H. Q. Wang, "Multi-fault rapid diagnosis for wind turbine gearbox using sparse Bayesian extreme learning machine," *IEEE Access*, vol. 7, pp. 773–781, 2019.

[5] W. G. Chen and H. L. Zhang, "Application of RF-LightGBM algorithm in early warning of wind turbine blade cracking," *Electronic Measurement Technology*, vol. 43, no. 1, pp. 162–168, 2020.

[6] G. L. Wang, H. S. Zhao, and Z. Q. Mi, "Application of XGBoost algorithm in prediction of wind motor main bearing fault," *Electric Power Automation Equipment*, vol. 39, no. 1, pp. 73–77, 2019.

[7] Z. D. Wu, X. L. Wang, and B. C. Jiang, "Fault diagnosis for wind turbines based on ReliefF and eXtreme gradient boosting," *Applied Sciences*, vol. 10, no. 9, 2020.

[8] J. Y. Hsu, Y. F. Wang, K. C. Lin, M. Y. Chen, and J. H. Y. Hsu, "Wind turbine fault diagnosis and predictive maintenance through statistical process control and machine learning," *IEEE Access*, vol. 8, pp. 23427–23439, 2020.

[9] S. Fan and Q. X. Tang, "Wind turbine pitch anomaly recognition system based on AdaBoost-SAMME," *Power System Protection and Control*, vol. 48, no. 21, pp. 31–40, 2020.

[10] S. B. Wang, X. G. Sun, and C. W. Li, "Wind turbine gearbox fault diagnosis method based on Riemannian manifold," *Mathematical Problems in Engineering*, Article ID 153656, 2014.

[11] D. M. Zhang, J. Yuan, J. Zhu, Q. Ji, X. Zhang, and H. Liu, "Fault diagnosis strategy for wind turbine generator based on the Gaussian process metamodel," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4295093, 2020.

[12] H. T. Chen, S. X. Jing, X. H. Wang, and Z. Y. Wang, "Fault diagnosis of wind turbine gearbox based on wavelet neural network," *Journal of Low Frequency Noise Vibration and Active Control*, vol. 37, no. 4, pp. 977–986, 2018.

[13] L. Lu, Y. G. He, T. Wang, T. C. Shi, and Y. Ruan, "Wind turbine planetary gearbox fault diagnosis based on self-powered wireless sensor and deep learning approach," *IEEE Access*, vol. 7, pp. 119430–119442, 2019.

[14] A. G. Kavaz and B. Barutcu, "Fault detection of wind turbine sensors using artificial neural networks," *Journal of Sensors*, vol. 2018, Article ID 5628429, 11 pages, 2018.

[15] B. D. Chen, P. J. Tavner, Y. H. Feng, W. W. Song, and Y. N. Qiu, *Bayesian Network for Wind Turbine Fault Diagnosis*, 2012, http://dro.dur.ac.uk/11029/1/11029.

[16] G. Y. Shi, J. Xu, and Q. Yang, "Intelligent fault diagnosis on wind turbine bearing based on convolutional neural network," *Journal of North China Electric Power University*, vol. 47, no. 4, pp. 71–79, 2020.

[17] X. L. Zhang, P. Han, L. Xu, F. Zhang, Y. Wang, and L. Gao, "Research on bearing fault diagnosis of wind turbine gearbox based on 1DCNN-PSO-SVM," *IEEE Access*, vol. 8, pp. 192248–192258, 2020.

[18] C. Zhang and C. B. Wen, "Fault detection of wind turbine blade based on improved Mask R-CNN," *Renewable Energy Resources*, vol. 38, no. 9, pp. 1181–1186, 2020.

[19] Y. H. Chang, J. L. Chen, C. Qu, and T. Y. Pan, "Intelligent fault diagnosis of wind turbines via a deep learning network using parallel convolution layers with multi-scale kernels," *Renewable Energy*, vol. 153, pp. 205–213, 2020.

[20] G. Q. Jiang, H. B. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2019.

[21] S. Yin, G. L. Hou, X. D. Yu, N. Li, Q. Wang, and L. J. Gong, "Research on temperature early warning method of wind turbine main bearing based on Bi-RNN," *Journal of Zhengzhou University*, vol. 40, no. 5, pp. 45–51, 2019.

[22] A. J. Yin, Y. H. Yan, Z. Y. Zhang, C. Li, and R. V. Sanchez, "Fault diagnosis of wind turbine gearbox based on the optimized LSTM neural network with cosine loss," *Sensors*, vol. 20, no. 8, p. 2339, 2020.

[23] H. F. Zheng, Y. R. Dai, and Y. Q. Zhou, "Fault prediction of fan gearbox based on K-means clustering and LSTM," *Materials Science and Engineering*, vol. 631, no. 2, pp. 32–43, 2019.

[24] H. F. Zheng and Y. R. Dai, "Fault prediction of fan gearbox based on deep belief network," *Journal of Physics*, vol. 1449, no. 1, 2020.

[25] J. H. Lei, C. Liu, and D. X. Jiang, "Fault diagnosis of wind turbine based on long short-term memory networks," *Renewable Energy*, vol. 133, pp. 422–432, 2019.

[26] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, Computer Science, 2014.

[27] Q. S. Liu, F. Zhou, R. L. Hang, and X. T. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sensing*, vol. 9, no. 12, pp. 1330–1335, 2017.

[28] J. R. Zhang, F. A. Liu, W. Z. Xu, and H. Yu, "Feature fusion text classification model combining CNN and BiGRU with multi-attention mechanism," *Future Internet*, vol. 11, no. 11, pp. 237–249, 2019.

[29] H. Y. Lv and Q. Feng, "A review of random forests algorithm," *Journal of Hebei Academy of Sciences*, vol. 36, no. 3, pp. 37–41, 2019.

[30] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423–1447, 2002.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6625758, 15 pages
https://doi.org/10.1155/2021/6625758

Research Article

A Data-Driven and Knowledge-Driven Method towards the IRP of Modern Logistics

Tiexin Wang,[1,2] **Yi Wu**,[1] **Jacques Lamothe**,[3] **Frederick Benaben**,[3] **Ruofan Wang**,[1] **and Wenjing Liu**[1]

[1]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*
[2]*Key Laboratory of Safety-Critical Software (Nanjing University of Aeronautics and Astronautics), Ministry of Industry and Information Technology, Nanjing 211106, China*
[3]*Centre Génie Industriel, IMT Mines Albi, Université de Toulouse, Albi, France*

Correspondence should be addressed to Tiexin Wang; tiexin.wang@nuaa.edu.cn

Received 31 December 2020; Revised 3 February 2021; Accepted 21 February 2021; Published 11 March 2021

Academic Editor: Xingsi Xue

Inventory Routing Problem (IRP) is a typical optimization problem in logistics. To reduce the total cost, which contains the product transportation cost, the inventory holding cost, the customer satisfaction cost, etc., a wide range of impact factors have to be taken into consideration. Since more and more intelligent devices have been adopted in the management of modern logistics, the amount of the collected data (relevant to those impact factors) increases exponentially. However, the quality of the collected data is suffering from a certain number of uncertainties, such as device status and the transmission network environment. Considering the volume and quality of the collected data, the traditional data-driven distribution optimization methods encounter a bottleneck. In this paper, we propose a hybrid optimization method which combines data-driven and knowledge-driven techniques together. In our method, a domain ontology, which has better scalability and generality, is built as an extension of data-driven optimization algorithms. Knowledge reasoning techniques are also combined to handle data quality issue and uncertainties. To evaluate the performance of our method, we carried out a case study, which is provided by a French company "Pierre Fabre Dermo-Cosmetics" (PFDC). This case study is a simplified scenario of the practical business process of PFDC.

1. Introduction

Supply chain is a network of organizations that are involved in collaborative processes that generate value as products and/or services in the end for the ultimate consumers [1]. Supply chain management [2] refers to the optimal supply chain operation, i.e., all activities from supply chain procurement to meeting the end customer at the lowest cost. Supply chain management contains several segments, such as supply chain strategy, supply chain planning, procurement, product life cycle management, and logistics. Supply chain management is aimed at integrating and coordinating the network [3].

Logistics, as one segment of supply chain management, refers specifically to the planning and implementing the flow of goods (or services). Traditionally, logistics is triggered by the inventory procurement. However, this kind of triggering mode has three main disadvantages [4]: (i) increase the burden of the enterprise investment, (ii) bear the risk of losing market opportunities, and (iii) force enterprises to engage in business activities which they are not good at. This results in a simple buy-to-sell relationship between suppliers and demand companies that does not solve some supply chain problems involving global strategic.

Today is an era of data explosion, in which every aspect of society is overwhelmed by the sheer volume of data generated [5]. Modern logistics, which can be regarded as a scenario of Internet of Things (IoT), generates and consumes a huge amount of data. Along a logistics, data is generated mainly from sensors, RFID, and production equipment. Furthermore,

data from other sources such as social media, newspapers, and weather forecast reports may also affect logistics. To help make efficient decisions and forecasts about logistics, leveraging these data with big data analytic techniques becomes a common practice for enterprises.

Towards the optimization of logistics, many data-driven methods (algorithms) such as "Vendor Management Inventory (VMI)" and "Collaborative Planning Forecasting and Replenishment (CPFR)" are proposed. As a typical issue of logistical optimization, the "Inventory Routing Problem (IRP)" attacks researchers' attention.

However, data-driven methods have inherent disadvantages. One of the typical disadvantages lies in handling uncertainties. For instance, a required data is missing or becomes incredible due to some unexpected reasons. In another instance, a decision is made by taking a set of impact factors into consideration, but some of the impact factors cannot be quantified (as computable data). Consequently, this kind of impact factors becomes uncertainties to data-driven methods. Therefore, concerning the optimization of modern logistics, three main challenges can be summarized as follows:

(i) Ch1: how to collect and make use of the heterogeneous data (on both syntax and semantic aspects) from different sources?

(ii) Ch2: concerning the data quality issue, how to ensure and improve the credibility of the collected data?

(iii) Ch3: for the impact factors that cannot be quantified and the uncertainties in logistics, how to measure and evaluate their influences?

Focusing on the three challenges, we propose a hybrid data-driven and knowledge-driven method for the optimization of modern logistics: DKDM4L. In DKDM4L, we adopt knowledge modelling and knowledge reasoning techniques to enhance data-driven methods. In the context of logistics, we formally define the relevant concepts, attributes, and relations by creating a domain ontology. In this domain ontology, concepts and attributes are defined with precise semantics, and constraints are added to the attribute values. By adopting knowledge reasoning techniques, the data incomplete issue and inconsistent issue are addressed. Furthermore, comparing to the pure data-driven methods, DKDM4L has stronger extendibility and generality. To evaluate the performance of DKDM4L, we carry out a practical use case, which is provided by PFDC. The main contributions of this work are as follows:

(i) Con1: in the context of logistics, we create an extensible domain ontology to formally describe domain concepts, attributes, and relations among them

(ii) Con2: by defining and applying knowledge reasoning rules on this domain ontology, we measure and evaluate the influence of uncertainties (e.g., weather conditions)

(iii) Con3: by cooperating with PFDC, we propose a practical use case (scenario) of modern logistics, which can be used as a baseline in this domain

The structure of this paper is as follows. The second section presents the motivated case provided by PFDC. The third section shows an overview of DKDM4L. The case study with evaluation is given in the fourth section. The fifth section illustrates the related works while a conclusion is given in the sixth section.

2. Motivated Case

2.1. Project Origin. The research work presented in this paper was initially triggered and founded by the European Horizontal 2020 project: Cloud Collaborative Manufacturing Network (C2Net). C2Net is aimed at providing a scalable real-time architecture, platform, and software to the supply network partners. The potential users of the C2Net platform are the small and medium-sized enterprises (SMEs), which do not currently have access to advanced management systems and collaborative tools due to their restricted resources.

Totally, there were around 20 partners taking part in this project. These partners came from both academics (research centers and laboratories) and industry (enterprises). C2Net had 7 work packages to cover the entire supply chain considering all stages of manufacturing, distribution, and sales. The research work presented in this paper originally belonged to work package 4, which focused on the optimization algorithms of logistics.

2.2. Practical Scenario. Pierre Fabre Dermo-Cosmetics (PFDC), as one partner of the C2Net Project, provided a practical scenario to simulate and evaluate logistics optimization algorithms.

PFDC is a French multinational pharmaceutical and cosmetic company. PFDC supply chain sources make and deliver products for a dermo-cosmetic market. PFDC manages 10 brands, more than 3500 product references, in around 140 countries over the world. PFDC supply chain concerns the following stakeholders: suppliers, manufacturing plants (in France), central distribution centers (in France), local subsidiaries or partners, and final customers (drugstores). Figure 1 shows a general overview of the PFDC business process.

In the local subsidiaries, local DRP (Distribution Requirement Planning) supported by FuturMaster solution is used to manage the forecasts and replenishments. In the central distribution center, central DRP (FuturMaster solution) and MRP II (Material Resources Planning) by SAP/ERP are used for the distribution and production planning.

2.3. Simplified Use Case. In order to focus on the distribution phase and simplify the real scenario, four hypotheses are defined in a simplified case, which is shown in Figure 2.

(i) H1: there is only one local distribution center (LDC), and it is in charge of delivering five kinds of products to five drug stores (DSs)

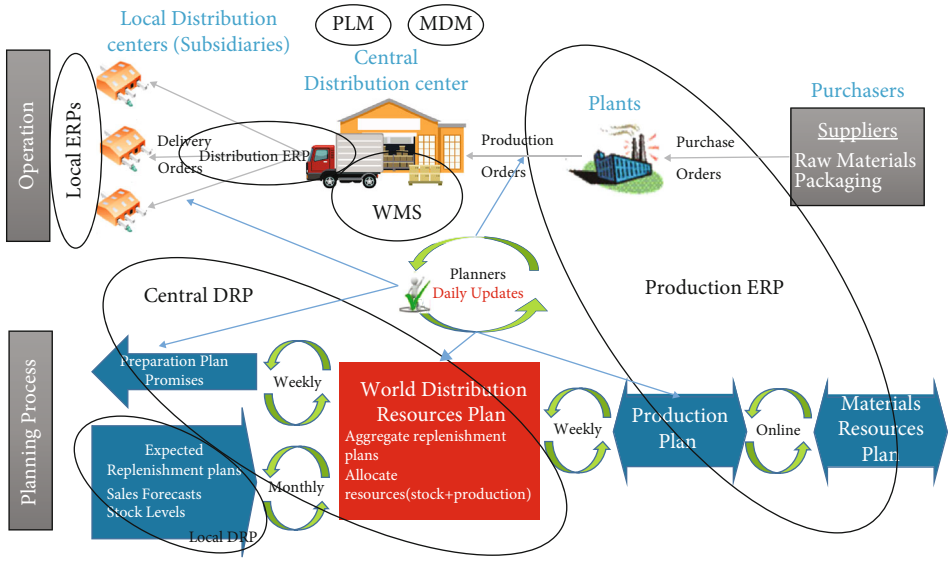


Figure 1: An overview of PFDC business process.

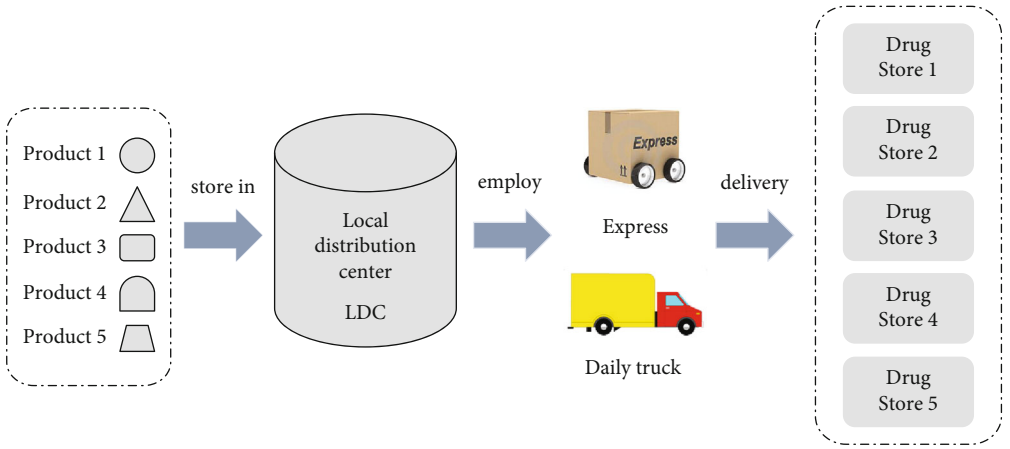


Figure 2: A simplified situation of the practical case.

(ii) H2: the unlimited manufacturing capacity, which means there are always enough products stored in the LDC. The inventory holding cost in LDC is ignored

(iii) H3: for each of the five DSs, the LDC sets two thresholds as the minimum inventory and the maximum inventory for each kind of products. These thresholds are key factors while making delivery decisions

(iv) H4: two kinds of delivery modes with different transportation costs can be used

(v) Express: it is the fastest one (one day lead time), and a one-to-one (the LDC to one specific DS) service. It is expensive, light load, a limit-number kind of products, etc.

(vi) Daily truck: it is a regular delivery mode, which has two days lead time. Meanwhile, it is a one-to-several (the LDC to several DSs). It is cheap, and the distribution route is fixed

For PFDC, the main goal of managing supply chain is to improve customer satisfaction. This means, at any time, stockouts are strictly prohibited in each of the five DSs' warehouses for all five kinds of products. On the other hand, considering the limited storage space and inventory holding costs, the number for all five kinds of products has an upper limit. A brief illustration of the constraints is shown in Figure 3.

There are several factors that are needed to be considered while setting the minimum and maximum thresholds. These factors concern both internal data and external data of PFDC. The internal data is always structured, such as retailers' sales reports, stock status, and historical sales. The external data can be both structured and unstructured, such as competing markets launch new products (from newspaper

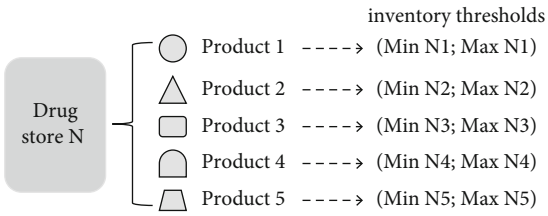


Figure 3: Managing DSs' inventory thresholds.

or video), new relevant guidance policies of government (from policy documents or TV). In order to set precisely specific thresholds for products in each DS, big data analytic techniques shall be used on all the data mentioned above.

2.4. IRP to Be Optimized. Towards IRP, the following three aspects of optimization have to be taken into consideration.

In order to reduce the costs of inventory holding in retailers (DSs), the inventory thresholds of various products stored in each DS should be set according to the sales situations. If in a specific period, the demand of one kind of products increases, the inventory thresholds should be raised to increase the delivery volume. Otherwise, the inventory thresholds should be lowered to reduce the delivery volume. Therefore, the thresholds of inventory shall be adjusted dynamically.

Delivery recommendation: DKDM4L suggests delivery plans for the inventory replenishment. Based on the sales forecasts of each DS, considering the required quantities (and volume) of all the products and the transportation costs, several potential delivery plans shall be made and recommended. The suggested plans concern on product packaging (quantities and weights), the transportation mode, and the distribution time.

Delivery route recommendation: if the daily truck transportation mode is triggered, a route planning is required for the truck. This recommendation concerns the optimization mainly on time and gas costs. Comparing to the former two issues, the route recommendation is not vital, and we do not take it into consideration in this paper.

3. Main Work

3.1. An Overview of DKDM4L. Considering the simplified use case provided by PFDC, we design a framework for DKDM4L. As shown in Figure 4, this framework contains four layers: (i) the physical layer that contains diverse sensors (e.g., from the drug stores, the LDC, and transportation vehicles), production machines, and IT instruments (e.g., servers and PCs), (ii) the data layer that is in charge of collecting and merging data, (iii) the model layer, which can be regarded as a knowledge model setting the unified syntax and semantics constraints of the collected data, and (iv) the reasoning layer, which defines the reasoning rules. The rules can be separated into two groups: one group to check (and correct) the consistency, integrity, and correctness of the collected data and another group of rules used to deduce the optimization distribution plans.

The first two layers mainly focus on data collecting, analyzing, and merging, while the last two layers concern more on the knowledge representing and reasoning. Therefore, DKDM4L is both a data-driven and knowledge-driven method.

First, the physical layer transmits the collected data to the data layer. Then, the data layer analyzes, validates, and transforms the received data to the unified forms and patterns that are defined in the model layer. Next, the model layer adds the formal semantics and relations to those well-prepared data. A large number of (knowledge) triples are generated on this layer. Finally, the reasoning layer uses these prepared triples to deduce the delivery decisions (optimized distribution and scheduling plans). Four layers, from bottom to top, are layer-by-layer dependent. The implementation of upper-layer functions relies on the services provided by its lower layer. Furthermore, the verification mechanism follows a top-down sequence.

The objective of designing this architecture is to improve the whole performance of supply relationship management. By separating different layers, staffs working on specific positions can focus only on their own roles. The data (and information) transition and verification between different layers shall be done automatically with mature protocols and software tools. This architecture supports the implementation of DKDM4L.

In the simplified use case, the product distribution happens only between a local distribution center (LDC) and five drug stores. The physical layer contains mainly the IT instruments (e.g., PCs, servers, intelligent sensors, and RFID) in the five drug stores, in the LDC, and on the transportation vehicles. In this paper, we focus mainly on the data layer, the model layer, the reasoning layer, and the connections among these three layers. By combining these three layers, the target of DKDM4L "calculating distribution plans that can reduce the total cost containing transportation cost, inventory holding cost and customer satisfaction cost, etc." can be achieved. The following three subsections present the details of the three layers, respectively.

3.2. The Data Layer. Nowadays, a huge amount of data is being generated at a high speed. IoT devices have been employed to provide new opportunities for sensing-based ubiquitous recognition and communication capabilities. However, since the diverse data sources and heterogeneous data (structured data, half-structured data, and unstructured data), making good use of these data becomes a tough task.

In the data layer of DKDM4L, there are four main data sources: data collected from LDC, data collected from retail (drug) stores, data collected from transportation vehicles, and the weather data. Table 1 is a general illustration of these data.

This table shows the collected data sources and the ways of collecting data. "Irrelevance" means the discrete data that are observed by sensors or manually input into computers, which is unrelated to each other. "Relevance" means that the value of the data is calculated based on other data rather than collecting from the direct data sources.

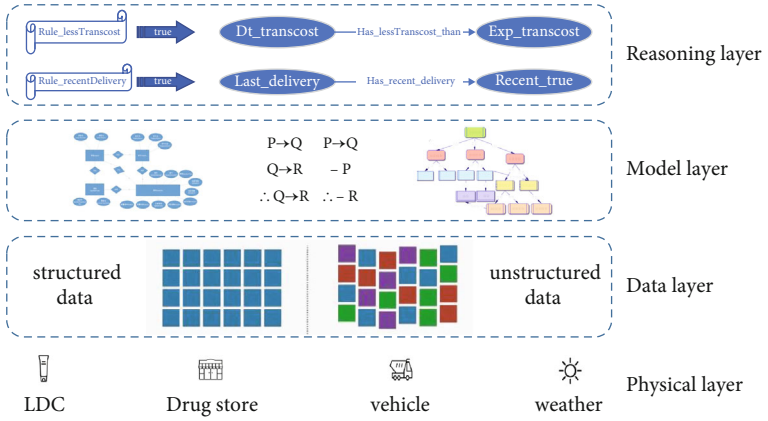


Figure 4: The architecture of DKDM4L.

Table 1: The collected data from the data layer.

Data source	Sensor collection	Computer records	Relevance		Data type	
			Relevant	Irrelevant	Structured	Unstructured
LDC		The product categories			√	
		Product inventory		√	√	
		Supply the drugstore			√	
Drug store		The product categories			√	
		Product inventories		√	√	
		Daily sales			√	
		Replenishment orders	√			√
Delivery center		Daily trunk number			√	
		Daily trunk cost		√	√	
		Express number			√	
		Express cost	√		√	
		Delivery route plan		√		√
		Delivery fee			√	
Delivery vehicles	Daily truck position					√
	Daily truck fault			√		
	Express position				√	
	Express the damage					
Weather	Sunshine intensity					
	Ultraviolet intensity					
	Rainfall intensity			√	√	
	Humidity					
	Haze					
	Time stamps					

Concerning the data about the weather, we partially employed "Roussey Catherine's weather ontology" [6]. Roussey Catherine describes a new meteorological dataset based on the SOSA/SSN ontology. This work is the first to publish meteorological data with the new version of the SOSA/SSN ontology. The network of the ontologies in [6] is composed of the following:

(i) Ontology to describe the different types of sensors

(ii) Ontology to describe the units of measurement

(iii) Ontologies to describe the geographical places and their locations

(iv) Ontology to describe the temporal entities

For example, the rain collector measures the quantity of precipitation that falls during a time period. The property "ssn:phenomenonTime" links the "sosa:Observation"

instance to an instance of the class "time:Interval." The properties "time:hasBegining," "time:hasEnd," and "time:hasDuration" specify the beginning, the end, and the duration of the interval, respectively.

The ontology describing the different types of sensors and the ontology describing the units of measurement are employed in DKDM4L to identify data sources. In addition, the domain ontology proposed in this paper also contains some weather factors, such as sunshine intensity and ultraviolet intensity, which are defined in [6] and are related to the skincare products (considering PFDC business).

3.3. The Model Layer. The modern logistics system involves many objects (classes), such as the LDC, retail stores, transportation vehicles, delivery plans, and weather types. Each object may have a certain influence to the system. To represent and simulate the numerous requirements and scenarios, modelling is a widely accepted engineering technology, which can achieve the management of the system [7]. Furthermore, the model layer is also in charge of identifying threats and vulnerabilities in the physical world based on relations and attribute values. Thus, models play a functional role not only in helping people understand the systems being developed but also in the management and detection of systems.

The ontological model is suitable for describing the dynamic environment of the Internet of Things applications. Furthermore, this kind of models can monitor, learn, and adapt to abnormal situations. A predefined adaptive knowledge base, as parts of the ontological model, can alert threats existing in the Internet of Things scenarios.

The definition of ontology [8, 9] contains four meanings: (i) conceptualizing domain knowledge, (ii) the concepts should be clear and unambiguous, (iii) formalizing the concepts, and (iv) the concepts should be good for sharing.

Ontology is defined as a five-tuple [10]: (i) concept; (ii) relationships—concepts are not isolated, they are interrelated; (iii) axiom—rules of reasoning; (iv) function—the mapping relationships between concepts; and (v) instance—unit objects that cannot be redivided.

Some of the existing mainstream knowledge representation languages are RDF [11], OWL [12], KIF [13], CycL [14], and OIL [15]. There are several main knowledge representing methods, such as the logical representation, the production representation, the frame representation, the object-oriented representation, the semantic web representation, the XML-based representation, and the ontology representation.

On the model layer, we propose a domain knowledge model "Inventory Routing Problem Ontology": IRPO. IRPO contains six aspects: the LDC knowledge representation, the drugstore representation, the transportation vehicle knowledge representation, the delivery knowledge representation, the weather knowledge representation, and the product knowledge representation. IRPO is defined conforming to the OWL 2.0 standard. Part of the formal representation is defined as follows:

<LDC, property, function, axiom, instance>
ObjectProperty: {has_product, has_drugstore}
<delivery, property, function, axiom, instance >
ObjectProperty: {delivered_To, has_ItemInfo}
DataProperty:{arrivalData:integer, delivery:interger, transportType:string}
<drugstore, property, function, axiom, instance>
DataProperty:{address:string, postalCode:string,}
ObjectProperty: {has_Delivery, has_Product, has_Transport}
Instance:{drugstore1, drugstore2, drugstore3, drugstore4, drugstore5}
<Inventory, property, function, axiom, instance>
Instance:{DS1_P1_Inventory, DS1_P2_Inventory, DS1_P3_Inventory}
ObjectProperty: {has_Inventory}
<iteminfo, property, function, axiom, instance>
ObjectProperty: {selling_Product}
DataProperty:{num:integer}
<cost,property,function,axiom,instance>
Subclass:{extracost, transcost}
<extracost, property,function,axiom,instance >
ObjectProperty: {is_ExtraCost}
DataProperty:{actualDemand:integer}
< transcost, property,function,axiom,instance >
DataProperty:{maxWeight:double, minWeight:double,-transPrice:double, DataProperty:{actualDemand:integer}}
<packageType, property,function,axiom,instance >
DataProperty:{weight:double}
Instance:{p1_Box, P1_Pack, P1_Pallet, P1_Unit}
<price, property,function,axiom,instance >
Instance:{DS1_P1,DS1_P2,DS1_P3, DS1_P4, DS1_P5}
DataProperty:{selling:double}
<product, property,function,axiom,instance >
Instance:{p1,p2,p3,p4,p5}
ObjectProperty: {has_ExtraCost, has_Inventory, has_Price}
DataProperty:{preparationCost:double, productType:string}
<transport, property,function,axiom,instance >
ObjectProperty: {has_TransCost}
DataProperty:{actualWeight:double, delay; integer, TransportType:string}
<weather, property,function,axiom,instance >
SubClass:{Sunshineintensity,Ultravioletintensity, Rainfallintensity,humidity,haze}

IRPO is constructed using Protégé. The structure of IRPO is shown in Figure 5. "*owl: Thing*" is superclass, which includes five modules "*Organization*," "*DeliveryCost*," "*DeliveryModel*," "*Weather*," and "*DailySelling*." "*Organization*" is the organizer who can include other supply modes by means of extending. "*Organization*" has three instances the "*LDC*," the "*Drugstore*," and the "*DeliveryCenter*". The "*LDC*" maintains "*Product*" and "*Inventory*"; each "*Product*" has "*Inventory*" as an attribute. Similarly, "*Drugstore*" also has "*Product*" and "*Inventory*" as attributes. A matter of concern in "*Drugstore*" is "*DailySelling*," which is affected by the "*Weather*." The "*Weather*" has five instances "*Shineintensity*," "*DeliveryModel*," "*Haze*," "*Humidity*," and "*Ult_inensity*." "*DeliveryModel*" has two instances "*Express*" and "*Daily_Truck*" as delivery modes. Both "*Delivery_Plan*" and "*DeliveryModel*" are scheduled by the "*DeliveryCenter*."

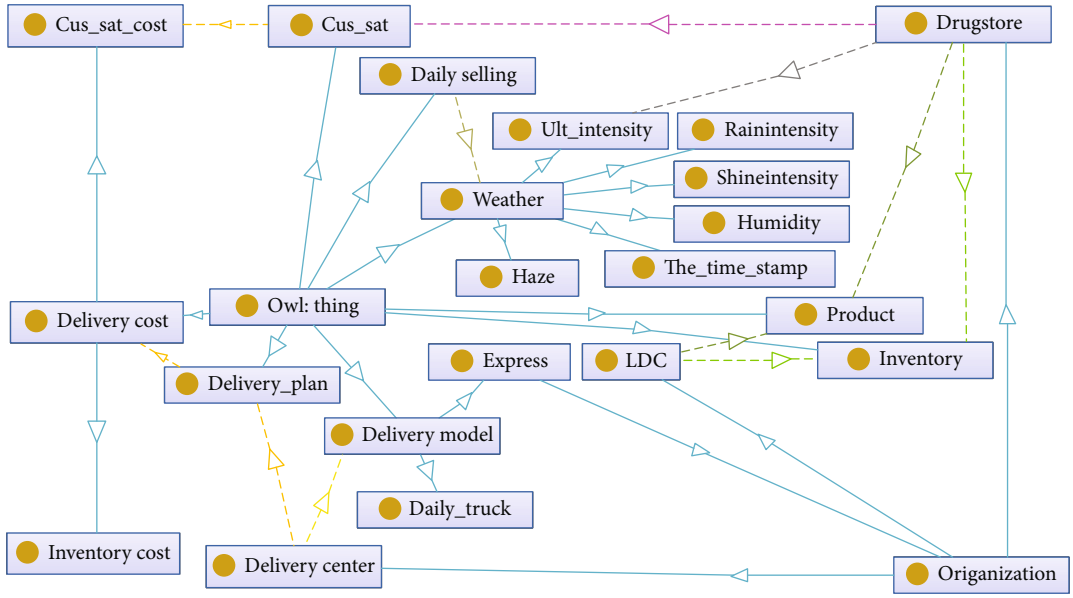


Figure 5: The domain ontology model.

Table 2: Relations employed in the domain ontology.

Relationship	Explanation	Instance
Extension	A class is a subclass of another class.	Owl:thing←(drugstore, LDC, delivery center) deliveryMode←(daily_truck, express)
Aggregation	A class consists of more than one class.	DeliveryCost←(transportcost, Inventorycost, customer_satisfication_cost)
Dependency	A class needs to use another class.	Dailysell is affected by the weather.

Additionally, "*Delivery_Plan*" concerns "*DeliveryCost*," while "*DeliveryCost*" consists of "*TransportCost*," "*InventoryCost*," and "*Cus_Sat_Cost* (*Custormer_Satisfaction_Cost*)." The "*Cus_Sat_Cost*" is affected by "*Cus_Sat* (*Custormer_Satisfication*)," which concerns about the maintained products inventory levels.

In IRPO, there are three kinds of relations defined among concepts. Table 2 lists the three relations and gives explanations about each of them. For each relation, an example is given to illustrate it.

3.4. The Reasoning Layer. Price and tax management [16] has been considered as a new management technology after supply chain management and customer relationship management (CRM). Its main idea is that a company should optimize the prices of its products and services based on a full understanding of the costs of the supply chain. At the same time, supply chain operations should also be optimized to reflect the revenue generated by different product types and customers. Therefore, prices and supply chain decisions should not be as independent as in the past but should be well integrated, which is another way to inject intelligence into supply chain management. Therefore, we propose the following three questions:

Q1: according to the sales of each retail (drug) store, how does the LDC distribute the product quantity?

Q2: according to distribution tasks, how to plan out a route with the lowest distribution cost?

Q3: for a retail (drug) store, considering the sales that are affected by weather factors, how to dynamically adjust the quantity of products delivered?

3.4.1. A Mathematical Model of Distribution Algorithms. The problem in the motivated case considers a local (subsidiary) distribution center that is in charge of delivering a set of products ($pi \in P$) to customer (drug stores) warehouses ($i \in W$) on a finite horizon ($t \in T$) through a VMI process. Customers are independent but admit to share the visibility on their demands ($d_{i,p}^{t}$). According to a CPFR midterm process, promises ($Pr_{i,p}^{t}$) of product availability have been made to each customer. Moreover, the current visibility of final consumer demands may no more fit to the one planned at the CPFR time. The problem thus arises when all the promises or all the demands cannot be fulfilled because of insufficient supply or production (R_{p}^{t}) at the local distribution center. From the vendor's perspective, the problem is to share the shortage ($\mathrm{IL}_{i,p}^{t}$) among the customers while avoiding stockouts ($I_{i,p}^{-t}$) at customer warehouses, satisfying as much as possible promises, synchronizing delivering tours, and respecting delivering frequencies (fr).

Various routes ($r \in R$) (i.e., multicustomer routes and emergency quick routes) have been defined beforehand for

delivering (parts of) the customers ($i \in \mathrm{Cr}$). This assumption dramatically reduces the IRP complexity, so that an optimization procedure can be used.

The underlying model is inspired from the [17] formulation but adds some specific constraints in order to model the supply limits and CPFR promised constraints. The model considers continue variables for inventories ($I_{i,p}^{t}$), transported quantities ($\mathrm{TR}_{i,p}^{t,r}$), low-level inventories ($\mathrm{IL}_{i,p}^{t}$), stockouts ($I_{i,p}^{-t}$), and nonsatisfied promises ($\mathrm{NPr}_{i,p}^{t}$). Integer variables formalize the decision of launching a transport on a route on a given time (z_r^t).

The objective function minimizes the total cost which contains transportation costs, various warehouse costs associated with inventory holding, nonrespect of the VMI minimal inventory costs, stockout costs, and nonrespect of the CPFR promises costs.

$$\mathrm{MIN}\left(\sum_{t\in T}\sum_{r\in R} fc_r.z_r^t + \sum_{t\in T}\sum_{i\in N\cup\{0\}}\sum_{p\in Pi} h_{i,p}.I_{i,p}^{t} + hl_{i,p}.\mathrm{IL}_{i,p}^{t} + h_{i,p}^{-}.I_{i,p}^{-t} + hp_{i,p}.\mathrm{NPr}_{i,p}^{t}\right). \tag{1}$$

Five of the general constraints are listed below. Constraints (2)and (3) express the balance of flows at drug store warehouses and the local distribution center. Constraint (4) defines the stockouts. Constraint (5) expresses VMI low-level inventory. Constraint (6) models the nonrespect of the CPFR promised quantities.

$$I_{i,p}^{t} = I_{i,p}^{t-1} + \sum_{r\in R, t>lr} \mathrm{TR}_{i,p}^{t-l_r,r} - d_{i,p}^{t}, \quad \forall t \in T, i \in W, p \in P_i, \tag{2}$$

$$I_{0,p}^{t} = I_{0,p}^{t-1} + R_p^t - \sum_{r\in R, t>lr} TR_{i,p}^{t,r}, \quad \forall t \in T, p \in P_i, \tag{3}$$

$$I_{i,p}^{t} + I_{i,p}^{-t} \geq 0, \quad \forall t \in T, i \in W, p \in P_i, \tag{4}$$

$$I_{i,p}^{t} + \mathrm{IL}_{i,p}^{t} \geq \mathrm{MIN}_{i,p}, \quad \forall t \in T, i \in W, p \in P_i, \tag{5}$$

$$\mathrm{NPr}_{i,p}^{t} = \mathrm{NPr}_{i,p}^{t-1} - \sum_{r\in R, t>lr} \mathrm{TR}_{i,p}^{t-l_r,r} + Pr_{i,p}^{t}, \quad \forall t \in T, i \in W, p \in P_i. \tag{6}$$

Some of the unitary costs are hard to be quantified and balanced. Inventory holding costs (hc) and freight costs (fc) can easily be measured, but low-level inventory and promised nonsatisfaction costs are rarely defined in the agreement. Thus, they can be defined as compromises in comparison to the other costs. Moreover, from the vendor's perspective, all the customers are rarely equivalent. So, holding costs are adapted so that important customers are favoured.

In the CPFR context, a demand is sensible to promotions and other market effects. That volatility makes it difficult to forecast. Thus, getting data from the market and modelling its impact on the demand forecasts become a crucial issue.

The PFDC algorithm is built upon a mathematical model to calculate and obtain the optimal distribution plans. However, the mathematical model is not suitable for ontology modelling in the first place. The mathematical model only relies on numerical calculation and has no semantic functions. Considering the weather factors, it cannot be dynamically programmed, so a more appropriate distribution optimization method is needed. In the field of knowledge engineering, rules are an important means to achieve reasoning [18].

The sales of cosmetics (PFDC products) are closely related to the weather, which can be divided into the following situations. When it is cloudy and rainy, the sunshine will weaken and the humidity will increase, which will restrain consumers from buying moisturizers and sunscreens. Sales of sunscreens and hydrating skincare products increase during the uV-heavy months. Promotional activities held to stimulate consumer consumption should also consider weather conditions to determine promotional products. In addition, studies have shown that consumers of combined products have higher sales than those of single products. Therefore, when considering combined products, weather factors should be taken into consideration to combine suitable products together, such as high-temperature weather, vigorous cleansing facial cleanser, and refreshing hydrating facial masks can be combined products.

3.4.2. The Definition and Division of Reasoning Rules. Reasoning refers to the process of introducing conclusions from existing facts according to certain rules. Knowledge-based reasoning rules emphasize the choices and applications of knowledge.

By adding semantic information to entities, semantic reasoning can be carried out to better realize the use of information. Ontological reasoning, with semantics as a prerequisite, can be automated by machines instead of manual reasoning. The following five types: (i) class hierarchy relationships, (ii) class equivalents, (iii) individual identity, (iv) compatible, and (v) classification, can be deduced automatically. The important role of ontological language in supporting reasoning includes checking the compatibleness of ontology and information, checking the implicit relations between classes, and automating the classification of instances. Automatic reasoning can check more content than manual reasoning, which is very beneficial to the large-scale ontology design or the fusion and sharing of data from different sources.

Ontological reasoning machines can be divided into two categories: special and universal. For the special ontology reasoning machines, some examples are Racer, FaCT, Pellet, etc. They support the main ontological languages, such as RDFS and OWL. For the universal ontological reasoning machines, one typical instance is Jess. At present, there are four main ways to implement ontological reasoning.

First, the reasoning methods are based on traditional description logic. Typical ones are Pellet [19], Racer [20], and FaCT [21], which are ontological reasoning machines designed and implemented based on traditional tableaux algorithms. Furthermore, many tableaux algorithm optimization techniques have been introduced to make efficient reasoning.

Second, the reasoning methods are based on rule-based approaches. Ontological reasoning, as a kind of application, can be mapped to the rule reasoning engine for reasoning.

Table 3: Rule types and source code.

Rule type	Rule	Source code
Distribution optimization	rule_delay rule_xsdailytrunc rule_xsexpress rule_recentDelivery Rule_badWeather	[(?d has_date ?x) (?x has_badWeather ?y) -> (?dhas_delayed_warning ?x)] [(?x possible_has_daily_trunc ?y) (?y equals_to_daily_trunc ?z) -> (?x actual_has_daily_trunc ?z)] [(?x possible_has_express ?y) (?y equals_to_express ?z) ->(?x actual_has_express ?z)] [(?x get_drug ?y) (?y equals_to_recent ?z) ->(?x has_recent_delivery ?z)] [(?x has_weather_infuence ?y)(?y equals_to_influence ?z) ->(?x has_badWeather ?z)]
Retail store inventory optimization	rule_storage rule_sellingall	[(?x reach_maxWarningNum ?y)(?y is_equal_to_notice ?z) ->(?x has_storage_warning ?z)] [(?x is_less_zero ?y) (?y is_equal_to ?z) -> (?x has_sellall_warning ?z)]
Cost optimization	rule_lessTranscost rule_fdailytrunc rule_fexpress	[(?x has_transcost ?y)(?y is_cheaper_than ?z) ->(?x has_lessTranscost_than ?z)] [(?x possible_take_daily_trunc ?y) (?y equals_to_dtt ?z) -> (?x take_daily_trunc ?z)] [(?x possible_take_express ?y) (?y equals_to_ext ?z) ->(?x take_express ?z)]

There are many ready-made conversion tools to implement OWL as reasoning rules. The ontological reasoning machines currently implemented as rule-based ones are Jess [22], Jena, etc.

Third, the reasoning methods are based on program editing. Based on the implementation of the deductive database technologies, two typical system projects are F-OWL and KAON2.

Fourth, the methods are based on the first-order predicate prover. Because OWL declaration statements can be easily converted into first-order logic, it is easy to use traditional first-order predicate provers to implement ontological reasoning for OWL, such as Hoolet's ontological reasoning machine, which uses Vampire's first-order predicate prover to implement ontological reasoning.

3.4.3. The Reasoning Rules Adopted in DKDM4L. We use SWRL [23] rule language to define the reasoning rules in DKDM4L. SWRL (Semantic Web Rule Language) is a language that renders rules semantically. Parts of the concepts of SWRL's rules are evolved by RuleML and combined with OWL ontology. SWRL is already a member of the W3C specification. SWRL can be regarded as a combination of rules and ontology. Through the combination of the two, the relationships and vocabulary depicted in ontology can be used directly while writing rules. While the relationships between these categories may otherwise require additional legal descriptions, ontology descriptions can be used directly in SWRL.

A total of ten rules with explanations are established as follows. Table 3 categorizes these ten rules into three groups and shows the source code.

(i) ***rule_delay***: *a delivery delay warning, if the delivery date spans a date with inclement weather on that date, the delivery may be delayed*

(ii) ***rule_storage***: *it is recommended to modify the stock prompt*

(iii) ***rule_sellingall***: *the inventory is less than or equal to zero inventory*

(iv) ***rule_xsdailytrunc***: *judgment may be dailyTrunc distribution, whether xs_dailyTrunc is true, if it is true can be dailyTrunc distribution or express delivery*

(v) ***rule_xsexpress***: *judgment for express delivery, whether xs_express to true, express delivery if said is true*

(vi) ***rule_lessTranscost***: *judging dailytrunc and express two distribution modes which cost is lower, the choice of the what kind of shipping method, if dailytrunc transport costs less than express transportation costs, transportation costs less equivalent to the total costs less, so choose dailytrunc, otherwise choose express*

(vii) ***rule_fdailytrunc***: *the final delivery way is dailytruck*

(viii) ***rule_fexpress***: *the final delivery way is express*

(ix) ***rule_recentDelivery***: *judge whether there is a recent distribution*

(x) ***Rule_badWeather***: *considers the day to be bad weather*

4. Case Study and Evaluation

4.1. Testing Data Acquired. In order to ensure the generality of the experimental case and make the performance of DKDM4L convincible, two aspects of effort have been made. First, the experimental case is generated from PFDC daily work; it is not just designed specifically for this research work. Second, all the testing data are real, which are captured from PFDC systems. To avoid obtaining the testing results by chance, we captured one-month period data and used all the

Table 4: Local distribution center, drug stores, and products.

Item	LDC	DS1	DS2	DS3	DS4	DS5	P1	P2	P3	P4	P5
Name	Muret	09260	11333	14311	19894	20307	511	691	585	677	686

Table 5: The prices of each product selling to different DSs.

	P1	P2	P3	P4	P5
DS1	4.84 €	4.66 €	10.21 €	4.47 €	8.12 €
DS2	4.47 €	4.54 €	10.21 €	4.47 €	8.12 €
DS3	4.84 €	4.54 €	10.29 €	4.47 €	8.12 €
DS4	4.47 €	3.89 €	10.21 €	4.47 €	8.12 €
DS5	4.84 €	4.54 €	10.21 €	4.47 €	8.12 €

data as testing input. A research team in PFDC provided the data captured from their daily business.

Parts of the sensitive records are replaced by particular items. A specific string of numbers is used to replace the real names of both drug stores and products. Table 4 shows the information of the LDC, drug stores, and products.

Due to some reasons, such as cooperation relations and purchase quantities, PFDC sells products to different DSs with different prices. The selling prices directly affect the inventory holding costs in DSs. Table 5 shows the selling prices of all products to each DS.

The weight and the volume of one product are two key issues to consider when making delivery plans. In the simplified situation, only the weights of each product are taken into consideration. The total weight directly affects the load and the cost of each distribution mode (especially the express mode). Table 6 shows the weights of each product. Here, only the net unit weights, which are provided by the producers of these products, are recorded.

As illustrated above, two kinds of distribution modes have been employed by PFDC. Each of them has floating costs based on both transportation distances and loads (weights) being carried. Table 7 shows the express distribution costs (concerning only the weights being distributed), and Table 8 shows the floating costs of the daily truck distribution mode. As shown in Table 7, the express distribution is used to deliver light loads and the cost increases with each kilogram.

If the delivery weight is more than ten kilograms, the express distribution is not a preferable mode due to its high costs. As shown in Table 8, the daily truck distribution is used to deliver medium loads; its cost increases with every ten kilograms. If the delivery weight is more than one hundred kilos, the distribution cost increases by 27.91 Euros per one hundred kilograms. Normally, this distribution mode covers all five DSs in one route.

Together with the transportation costs, the inventory level is another important driving factor in making distribution decisions. As illustrated in Figure 3, two thresholds are defined to limit the inventory for all five kinds of products in each DS. Table 9 shows all the threshold pairs.

Considering the distribution driving factors from the inventory aspect, besides these threshold pairs, there are

Table 6: The weights of each product.

	P1	P2	P3	P4	P5
Unit/weight (kg)	0.063	0.369	0.485	0.056	0.122

two other items "current inventory records" and "daily sell outs." The current inventory records and daily sell outs are real-time data that are automatically generated. PFDC provided the current inventory records (CIR) and daily sell out (DSO) data collected within a period of one month. As an example, Table 10 shows parts of the data collected from drug store 2 during a five-day period.

4.2. The Testing Results

4.2.1. The Testing Process. The original collected data is one-month sale data of PFDC, including initial inventories, daily sales of each drug store. The original data has 5 stores (DS1, DS2, DS3, DS4, and DS5) with 5 items to sell (P1, P2, P3, P4, and P5). Using the original nearly one-month sale data as input, the daily inventories and sales of products and inventory restrictions can be obtained through model processing.

4.2.2. The Results and Evaluation. DKDM4L recommends the optimization distribution plans for PFDC, considering the delivery costs, the delivery time, the drug store inventory thresholds, etc.

Figure 6 shows the comparison results between original distribution plans and suggested distribution plans by DKDM4L. The horizontal coordinate indicates the date while the vertical coordinate represents the total costs of delivery. According to this chart, the direct distribution costs of DKDM4L suggested plans are slightly higher than the original plans. Particularly, on the first day, the distribution cost contributed almost one-third of the total costs. However, the original plans can not strictly satisfy the "minimum and maximum inventory" restriction, which threatens the stable supply of products. The two charts, shown in Figures 7 and 8, about "DS1-P2" and "DS4-P2" inventory changes briefly demonstrate this point.

As shown in Figures 7 and 8, the plans suggested by DKDM4L satisfied perfectly the "maximum and minimum" inventory restriction.

Actually, the distribution plans suggested by DKDM4L largely enhances the stability of product supply, benefiting the drug stores in the long term as well as earning them a good reputation. To PFDC, the customer satisfaction always comes first.

(i) Optimization of the product delivery. Input the one-day sales data of each product in each drug store, and DKDM4L gives the total amount of required products for the next day. DKDM4L can generate distribution alerts before (and after) the

Table 7: The costs of the express distribution mode.

To.\weight (kg)\cost (€)	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
DS1/DS4/DS5	3.93	4.41	5.38	5.84	6.31	6.77	7.23	7.64	8.05	8.46
DS2	3.86	4.33	5.30	5.76	6.22	6.68	7.12	7.53	7.93	8.34
DS3	3.81	4.22	4.99	5.40	5.81	6.22	6.44	6.66	6.89	7.11

Table 8: The costs of the daily truck distribution mode.

To.\weight (kg)\cost (€)	1-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	<100
DS1/DS4/DS5	13.14	14.98	16.73	18.70	20.60	23.04	25.10	26.62	28.13	28.99
DS2	13.42	15.30	17.08	19.10	21.04	23.53	25.63	27.18	28.72	29.60
DS3	11.13	12.79	14.24	15.80	17.42	19.32	21.03	22.40	23.55	24.51

Table 9: Inventory managing threshold pairs.

DS/product	P1		P2		P3		P4		P5	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
DS1	20	50	20	50	1	5	1	10	2	10
DS2	30	60	50	200	1	6	5	20	2	10
DS3	20	100	15	250	1	50	2	10	3	15
DS4	30	150	200	500	3	20	10	50	30	70
DS5	30	100	100	1000	10	100	5	30	10	50

Table 10: Current inventory records and daily sell out data collected from DS2.

Product/day	Day 1		Day 2		Day 3		Day 4		Day 5	
	CIR	DSO	CIR	DSO	CIR	DSO	CIR	DSO	CIR	DSO
P1	36	6	30	5	25	7	18	5	31	8
P2	35	8	27	9	18	10	8	8	28	0
P3	3	1	3	0	3	1	2	0	2	0
P4	21	1	19	1	18	2	16	3	13	0
P5	23	1	21	2	31	2	29	1	6	2

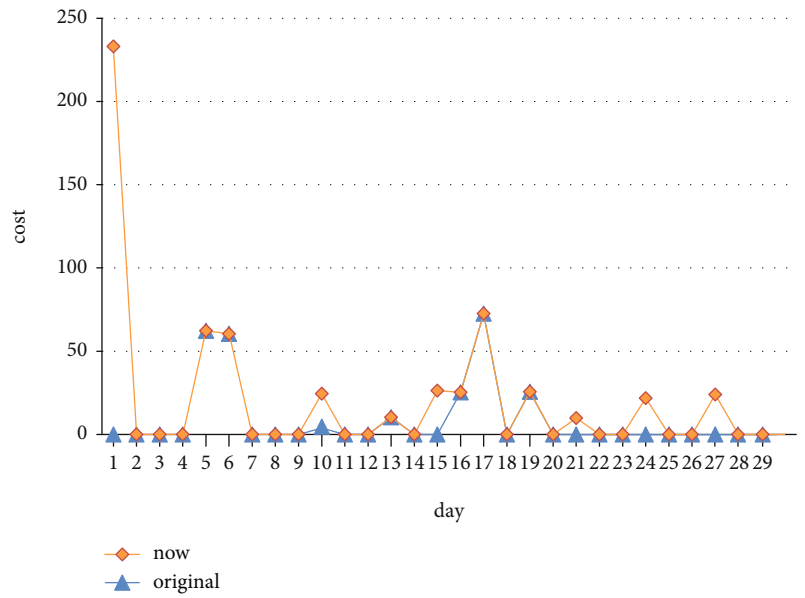


Figure 6: A comparison of total costs between original plans and DKDM4L suggested plans.

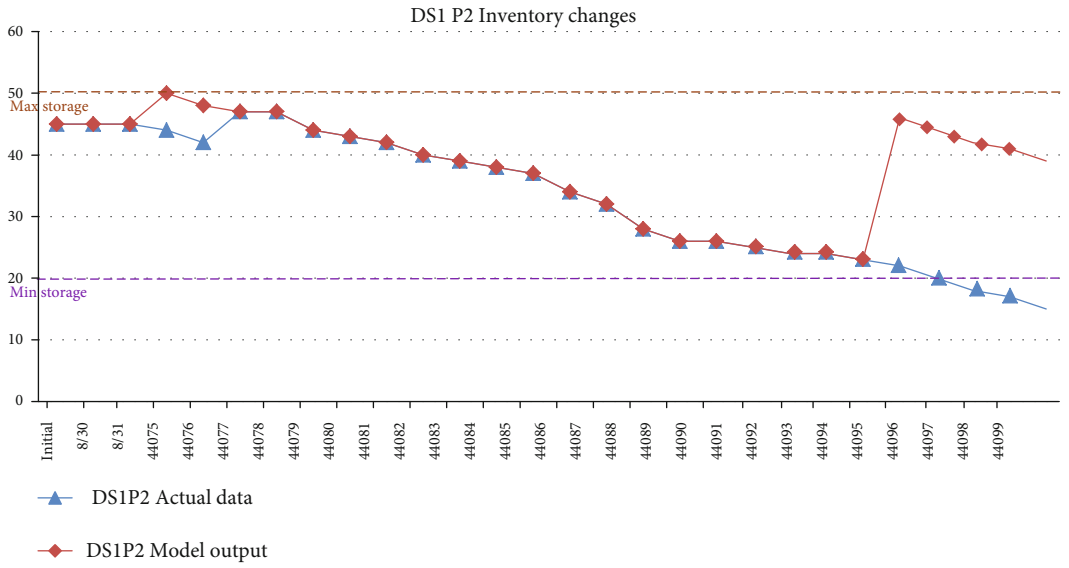


Figure 7: The inventory maintaining situation of P2 in DS1.

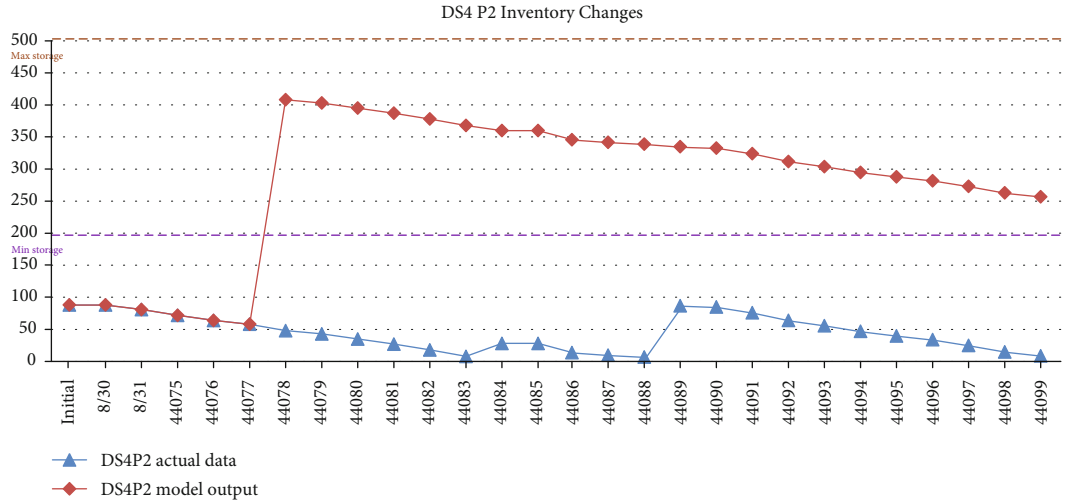


Figure 8: The inventories maintaining situations of P2 in DS4.

recommended plans being executed. Four kinds of alerts are listed as follows:

(a) The sales (decreased significantly) warnings

(b) Inventory (below the minimum inventory threshold) warnings

(c) The total number of noncompliance requirement warnings

(d) Inventory (above the maximum inventory threshold) warnings

(ii) Optimization of the delivery mode. After a delivery plan is formed, a recommendation of the delivery model (express or daily truck) is given based on the quantity and volume of the delivery products and the leading time of the delivery mode. DKDM4L considers also the weather influence that can cause delivery delays

(iii) Recommendation of the optimization of drug store inventory thresholds. DKDM4L takes the weather conditions into account, predicting the delivery arrival time. If the weather conditions may trigger a sold-out warning, both the minimum and maximum inventory thresholds will be raised, and vice versa

To sum up, DKDM4L has the following two advantages: ensure the continuous supply of products and bring customers (drug store) a better being served experience. One point to be emphasized, PFDC is a pharmaceutical and cosmetic company, and the selling of these kinds of products is

sensitive to the weather. DKDM4L is aimed at serving this kind of companies to optimize their logistics. This is also a limitation on the usage of DKDM4L.

5. Related Work

We present the related work from two aspects: traditional data-driven supply chain management (especially focusing on the logistics issue) methods and modern knowledge-driven optimization supply chain management methods.

For the traditional data-driven logistics management methods, "Collaborative Planning Forecasting and Replenishment" (CPFR) and "Vendor Managed Inventory" (VMI) are identified as mutually benefiting good practices [24, 25].

CPFR structures long to mid-term planning processes so that partners jointly plan a number of promotional activities and work out synchronized forecasts, on the basis of which the production and replenishment processes are determined [26]. VMI is based on an agreement where vendor and buyer agree on a process for sharing data (product sales, forecasts about future sales, and inventory levels) so that the vendor monitors the customers' inventory organizing replenishments (deciding order quantities, shipping, and timing) [27]. The vendor can take advantage of these data to dynamically adapt lot sizes, synchronize deliveries to several customers, and adjust the delivery frequency.

From the decision support point of view, VMI falls under the Inventory Routing Problem (IRP). Knowing a planned demand on some customers, the IRP objective is to decide on delivery quantities and maintain the customers' inventories in an agreed range while organizing distribution tours in order to minimize the total cost of the supply chain [28]. The complexity of the problem depends on the number of products, the horizon of decisions (1 period, finite or infinite horizon), the nature of demands (planned or stochastic), the existence of routing alternatives (exist or must be built), and the vendor constraints (finite quantities per product, finite or infinite production capacity). To solve this problem, many heuristics and optimization data-driven procedures have been proposed depending on the specificities of the problem [7]. To acquire good performance, both CPFR and VMI require the support of large quantity and high-quality relevant data. The process of collecting, storing, retrieving, and processing data is important to apply the two practices.

However, since the IoT theories and techniques become mature, more and more intelligent devices are employed in logistics. These devices generate a large volume of heterogonous data with a high speed. Meanwhile, considering the devices themselves, the network transmission environment, and data processing techniques, the quality of these data is difficult to ensure. Furthermore, some impact factors that cannot be quantified and certain uncertainties also affect the distribution decisions in logistics. Considering the above factors, the applications of data-driven optimization algorithms have encountered a bottleneck.

The advanced data transmission and storage technologies, such as wireless sensor networks (WSNs), enabled modern logistics. A large number of research works focusing on sensor data management are published. In [29], the authors focus on the technologies of optimizing the data storage of wireless sensors (WSNs); blockchain technology is introduced to save the storage space of network nodes. In [30], the authors focus on the data redundancy problem; a two-stage data simplicity method for the sensor network is proposed.

Even though data processing technologies are improved, uncertainty issues still cannot be handled well by data-driven optimization methods. Therefore, knowledge-driven methods have been proposed in both academics and industry. The adaptation of knowledge representation (with domain ontologies) and knowledge reasoning in IoT applications (e.g., supply chain management, smart home, and e-health) becomes quite common now. In the medical Internet of Things, which aims at realizing the local ontological semantic expansion by being associated with open correlation data sources, research work [31] proposed an ontology model that is designed and applied to multiple sensors to collect vital sign data. Since the concept of intelligent supply chain [32] was put forward, more and more researchers have paid attention on the combination of supply chain management and IoT. Research work [33] reviewed the applications of big data analysis technologies in supply chain management. In [34], the authors built the "TOVE Traceability Ontology" to trace the source of products. There are other domain ontologies built in the context of supply chain management, such as works presented in [35, 36]. Research works presented in [37, 38] are also knowledge-driven methods focusing on traceability of delivering products. Reference [39] focuses on configuring blockchain architectures for supply chain. In [40], a noteworthy effort develops the EAGLET ontology for ensuring data interoperability between diverse IoT devices over a supply chain.

Benefiting from the rapid development of information technologies, the cost of logistics has been greatly reduced. In order to make good use of those relevant technologies, domain-specific adaptation is required. Traditional data-driven distribution optimization algorithms have to be adapted and enhanced to face new challenges (e.g., uncertainties brought by big data era). Ontology, as a typical way of representing knowledge, has been adapted widely in the combination of supply chain management and IoT. Focusing on the specific IRP, this paper focuses mainly on improving the performance of mature data-driven optimization algorithms with knowledge-driven theories and techniques. Particularly, knowledge reasoning is introduced to handle uncertainties and factors that cannot be quantified.

6. Conclusion

This paper proposes a hybrid data-driven and knowledge-driven method "DKDM4L" to optimize the IRP of modern logistics. A four-layer theoretical framework is proposed, and as the core of this framework, specific domain ontology is created on the third layer. This domain ontology is built upon two mature optimization algorithms of IRP, and the mechanism of handling factors that cannot be quantified and uncertainties has been integrated in as functions and reasoning rules.

Compared to the traditional data-driven IRP optimization methods, DKDM4L owns three main advantages. First, based on the formal precise semantics of the domain ontology, DKDM4L can better handle data quality issues, such as believability and completeness. Second, as an inherent characteristic of knowledge-driven methods, DKDM4L has better scalability and generality. This means DKDM4L can be tailored or extended easily for other applications. Third, uncertainty (especially considering weather conditions) handling mechanism has been integrated in DKDM4L. With the editable reasoning rules defining on the fourth layer of the framework, the product distribution decisions made by DKDM4L are more reasonable. Based on the three advantages, DKDM4L can be a better potential solution to modern logistics, which can be regarded as a practical scenario of IoT.

The original trigger of this research work is the C2Net Project. At first, we focused only on proposing a new VMI optimization algorithm (a pure data-driven one). As one partner of the C2Net Project, PFDC provided a practical business scenario with real data as the test case. During the project, we found that a pure data-driven optimization method had encountered many limitations. Therefore, we extended our work to the current status. Again, the practical scenario (a simplified version with four hypotheses) from PFDC is the foundation of proposing DKDM4L. The performance of DKDM4L has also been tested and evaluated with the collected real data. The testing results approve that DKDM4L is a potential solution to IRP of modern logistics.

We will extend DKDM4L in three aspects in the future. First, enrich the domain ontology (knowledge model) to improve the generality. The current domain ontology is built mainly considering the scenario provided by PFDC. More scenarios and industry standards will be taken into consideration to enrich this ontology. Second, more uncertainty handling mechanisms will be included. Besides the influence of weather conditions, other uncertainties such as the launching of new products, the promotion of other competitive products that also affect on distribution decisions should be well addressed. Third, by analyzing industry standards, more reasoning rules are necessary to be defined both to better control the quality of collected data and better address other uncertainties.

Data Availability

The whole testing dataset provided by PFDC is available. We can share it with researchers providing a formal application.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2018YFB1003900), the Natural Science Foundation of China (61872182), and the CCF-Huawei Database System Innovation Research Plan under Grant CCF-HUAWEI DBIR2020001A. The authors also would like to acknowledge the financial support given by the C2Net Project as a trigger of this work.

References

[1] S. Salhi and M. Christopher, "Logistics and supply chain management: strategies for reducing costs and improving services," *Journal of the Operational Research Society*, vol. 45, no. 11, pp. 1341–1341, 1994.

[2] D. Saidi, J. El Alami, and M. Hlyal, "Sustainable supply chain management: review of triggers, challenges and conceptual framework," *IOP Conference Series: Materials Science and Engineering*, vol. 827, article 012054, 2020.

[3] H. Stadtler, "Supply chain management: an overview," in *Supply Chain Management and Advanced Planning*, H. Stadtler, C. Kilger, and H. Meyr, Eds., pp. 3–28, Springer, Berlin Heidelberg, 2015.

[4] L. Lu, J. Marín-Solano, and J. Navas, "An analysis of efficiency of time-consistent coordination mechanisms in a model of supply chain management," *European Journal of Operational Research*, vol. 279, no. 1, pp. 211–224, 2019.

[5] B. Saha and D. Srivastava, "Data quality: the other face of big data," in *2014 IEEE 30th International Conference on Data Engineering*, pp. 1294–1297, Chicago, IL, USA, March 2014.

[6] C. Roussey, S. Bernard, G. André, and D. Boffety, *Weather data publication on the LOD using SOSA/SSN ontology*, Semantic Web, 2019.

[7] L. C. Coelho, J. F. Cordeau, and G. Laporte, "Thirty years of inventory routing," *Transportation Science*, vol. 48, no. 1, pp. 1–19, 2014.

[8] M. Jarrar, J. Demey, and R. Meersman, "On using conceptual data modeling for ontology engineering," in *Journal on Data Semantics I*, pp. 185–207, Springer, Berlin, Heidelberg, 2003.

[9] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?," in *Handbook on Ontologies*, pp. 1–17, Springer, Berlin, Heidelberg, 2009.

[10] G. Harman, *Object-oriented ontology: A new theory of everything*, Penguin UK, 2018.

[11] E. Miller, "An introduction to the resource description framework," *Bulletin of the American Society for Information Science and Technology*, vol. 25, no. 1, pp. 15–19, 1998.

[12] S. Bechhofer, F. Van Harmelen, J. Hendler et al., "OWL web ontology language reference," *W3C Recommendation*, vol. 10, no. 2, 2004.

[13] A. S. Abdelghany, N. R. Darwish, and H. A. Hefni, "An agile methodology for ontology development," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 2, pp. 170–181, 2019.

[14] D. B. Lenat and R. V. Guha, "The evolution of CycL, the Cyc representation language," *ACM SIGART Bulletin*, vol. 2, no. 3, pp. 84–87, 1991.

[15] D. L. McGuinness, R. Fikes, J. Hendler, and L. A. Stein, "DAML+ OIL: an ontology language for the semantic web," *IEEE Intelligent Systems*, vol. 17, no. 5, pp. 72–80, 2002.

[16] A. Parssian, S. Sarkar, and V. S. Jacob, "Assessing data quality for information products: impact of selection, projection, and cartesian product," *Management Science*, vol. 50, no. 7, pp. 967–982, 2004.

[17] J. F. Cordeau, D. Laganà, R. Musmanno, and F. Vocaturo, "A decomposition-based heuristic for the multiple-product

inventory-routing problem," *Computers & Operations Research*, vol. 55, pp. 153–166, 2015.

[18] Y. L. Chi, "Rule-based ontological knowledge base for monitoring partners across supply networks," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1400–1407, 2010.

[19] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: a practical owl-dl reasoner," *Journal of Web Semantics*, vol. 5, no. 2, pp. 51–53, 2007.

[20] V. Haarslev and R. Möller, *Racer: a core inference engine for the semantic web*, EON, 2003.

[21] D. Tsarkov and I. Horrocks, "FaCT++ description logic reasoner: system description," in *Automated Reasoning*, pp. 292–297, Springer, Berlin, Heidelberg, 2006.

[22] E. J. Friedman-Hill, *Jess, The Rule Engine for the Java Platform*, 2008, http://herzberg.ca.sandia.gov/jess/.

[23] World Wide Web Consortium, "A semantic web rule language combining OWL and RuleML," 2004, http://www.w3.org/Submission/SWRL.

[24] D. Z. T. S. Z. Ming and Y. D. C. Jie, "Overview of ontology," *Acta Scicentiarum Naturalum Universitis Pekinesis*, vol. 5, no. 27, 2002.

[25] R. Akerkar and P. Sajja, *Knowledge-Based Systems*, Jones & Bartlett Publishers, 2009.

[26] F. Panahifar, C. Heavey, P. J. Byrne, and H. Fazlollahtabar, "A framework for collaborative planning, forecasting and replenishment (CPFR)," *Journal of Enterprise Information Management*, vol. 28, no. 6, pp. 838–871, 2015.

[27] P. Danese, "Towards a contingency theory of collaborative planning initiatives in supply networks," *International Journal of Production Research*, vol. 49, no. 4, pp. 1081–1103, 2011.

[28] N. H. Moin and S. Salhi, "Inventory routing problems: a logistical overview," *Journal of the Operational Research Society*, vol. 58, no. 9, pp. 1185–1194, 2007.

[29] Y. Ren, Y. Liu, S. Ji, A. K. Sangaiah, and J. Wang, "Incentive mechanism of data storage based on blockchain for wireless sensor networks," *Mobile Information Systems*, vol. 2018, Article ID 6874158, 10 pages, 2018.

[30] H. Harb, A. Makhoul, and C. Abou Jaoude, "A real-time massive data processing technique for densely distributed sensor networks," *IEEE Access*, vol. 6, pp. 56551–56561, 2018.

[31] M. G. Seneviratne, M. G. Kahn, and T. Hernandez-Boussard, "Merging heterogeneous clinical data to enable knowledge discovery," in *Biocomputing 2019*, pp. 439–443, Kohala Coast, HI, USA, January 2019.

[32] G. R. Lv and Q. Yu, "The construction of intelligent supply chain," in *Advanced Materials Research*, vol. 694, pp. 3567–3570, Trans Tech Publications Ltd., 2013.

[33] N. Agrawal and S. A. Smith, "Optimal inventory management using retail prepacks," *European Journal of Operational Research*, vol. 274, no. 2, pp. 531–544, 2019.

[34] H. M. Kim and M. Laskowski, "Toward an ontology-driven blockchain design for supply- chain provenance," *Intelligent Systems in Accounting, Finance and Management*, vol. 25, no. 1, pp. 18–27, 2018.

[35] D. Frey, P. O. Woelk, T. Stockheim, and R. Zimmermann, "Integrated multi-agent-based supply chain management," in *Twelfth IEEE international workshops on enabling technologies: infrastructure for collaborative enterprises*, pp. 24–29, Linz, Austria, June 2003.

[36] T. Grubic and I. S. Fan, "Supply chain ontology: review, analysis and synthesis," *Computers in Industry*, vol. 61, no. 8, pp. 776–786, 2010.

[37] A. Smirnov and C. Chandra, "Ontology-based knowledge management for co-operative supply chain configuration," in *The proceedings of the 2000 AAAI spring symposium "Bringing knowledge to business processes"*, pp. 85–92, Stanford, CA, USA, March 2000.

[38] D. E. O'Leary, "Supporting decisions in real-time enterprises: autonomic supply chain systems," in *Handbook on Decision Support Systems 2*, pp. 19–37, Springer, Berlin, Heidelberg, 2008.

[39] H. Ringsberg, *Traceability in Food Supply Chains Exploring Governmental Authority and Industrial Effects*, Department of Design Sciences, Faculty of Engineering, Lund University, 2011.

[40] D. E. O'Leary, "Configuring blockchain architectures for transaction information in blockchain consortiums: the case of accounting and supply chain systems," *Intelligent Systems in Accounting, Finance and Management*, vol. 24, no. 4, pp. 138–147, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6631074, 11 pages
https://doi.org/10.1155/2021/6631074

Research Article

Intelligent Recognition System Based on Contour Accentuation for Navigation Marks

Yanke Du[1], **Shuo Sun**[1], **Shi Qiu**[1], **Shaoxi Li**[1], **Mingyang Pan**[1] and **Chi-Hua Chen**[2,3]

[1]*Navigation College, Dalian Maritime University, Dalian 116026, China*
[2]*College of Mathematics and Computer Sciences, Fuzhou University, Fuzhou 350108, China*
[3]*Key Laboratory of Intelligent Metro of Universities in Fujian, Fuzhou University, Fuzhou 350108, China*

Correspondence should be addressed to Mingyang Pan; panmingyang@dlmu.edu.cn and Chi-Hua Chen; chihua0826@gmail.com

Received 27 December 2020; Revised 1 February 2021; Accepted 18 February 2021; Published 3 March 2021

Academic Editor: Xingsi Xue

Sensing navigational environment represented by navigation marks is an important task for unmanned ships and intelligent navigation systems, and the sensing can be performed by recognizing the images from a camera. In order to improve the image recognition accuracy, this paper combined a contour accentuation algorithm into a multiple scale attention mechanism-based classification model for navigation marks. Experimental results show that the method increases the accuracy of navigation mark classification from 95.98% to 96.53%. Based on the classification model, an intelligent navigation mark recognition system was developed for the Changjiang Nanjing Waterway Bureau, in which the model is deployed and updated by the TensorFlow Serving.

1. Introduction

For unmanned ships and vessel traffic service (VTS), intelligent perception of the navigational environment is an important topic [1]. The navigational environment mainly includes two parts: the dynamic vessels and the navigational features marked by the navigational aids. According to the IALA (International Navigation and Lighthouse Administration Navigation Association), the definition of the term Aid to Navigation (AtoN) means any specific equipment, system, or service outside the ship, specifically used to assist navigators in determining their location or safe route or to warn them of dangers or obstacles to navigation [2]. AtoN mainly consists of buoy and beacon; the former is a floating object fixed at the bottom; the latter is a structure permanently set on the seabed or land. Both of them can be categorized as "marks." They have distinctive shapes, colors, top marks, and other auxiliary markings which can be observed to indicate their purposes during the daytime. The relevant information about navigation marks is usually obtained through the Electronic Chart Display and Information System (ECDIS) or Automatic Identification System (AIS) [3]. However, it is a new challenge about how to visually and automatically detect navigational aids through the camera.

With the development of artificial intelligence technology, many intelligent detection technologies are applied to VTS [4, 5] and smart ships [6, 7]. Among them, the applications of deep learning technology in the detection and classification of ships are currently widely used [8, 9]. The purpose of this type of application is to supplement information about ships not in AIS [10]. For navigation marks, although their basic information can be obtained through ECDIS, sailors are still required to keep visual observing on their realtime states by the eye or with a telescope [11]. At the present stage, the detection and classification of navigation mark images is not as widely studied as ships [12]. Compared to ship image classification and recognition, there are fewer references available. In previous research [12], we exploited deep learning technology to study the navigation marks' image recognition during the daytime. It proposed a fine-grained ResNet-based classification model to classify navigation marks named ResNet-Multiscale-Attention (RMA). The accuracy

of this model reaches at 95.98% on a dataset including 10260 navigation mark images. However, the experimental results showed that the model also has some certain misclassifications of navigation marks, especially in aspect of the images with inconsistent shapes.

To solve these problems, this paper studied further to improve the classification model for navigation mark images, and the contributions are highlighted as follows.

(i) An improved navigation mark classification method with contour accentuation is proposed, and its classification accuracy arrives in 96.53%

(ii) An intelligent service system is developed and has been applied by the Changjiang Nanjing Waterway Bureau; it provides image recognition service of navigation marks on the Yangtze River

The contents of this article are organized as follows. Section 2 describes the related works. Section 3 describes the improved classification model for navigation mark images by contour accentuation method. Section 4 provides practical experimental results and discussion. Section 5 illustrates the intelligent application system. Finally, conclusions and future work are given in Section 6.

2. Related Work

In deep learning technology, convolution neural networks (CNN) are suitable for visual recognition and image classification tasks. AlexNet [13], VGG [14], GoogleNet [15], ResNet [16], and DenseNet [17] are some of the networks that attract attention from researchers. Various image classification methods based on CNN were applied to many fields, such as medical image analysis [18] and face recognition [19]. Some researches about vessel recognition also had been reported. Shi et al. [20] put forward a new deep learning framework, which combined the underlying functions and could effectively use useful information to classify the ship optical image. Oliveau et al. [21] proposed a new vessel classification theory based on semisupervised learning. Shin et al. [22] proposed a model using interest region combined a convolutional neural network for improving the ship images' classification accuracy. Solmaz et al. [23] proposed a framework and a new loss function to recognize the marine and land vehicles in a fine-grained way using multitasking learning.

Comparing with the vessel images, the different types of navigation marks may only have subtle differences in certain specific positions. To some extent, their image classification is a fine-grained classification. An important method of fine-grained classification is the attention mechanism. The attention mechanism is essentially to imitate the way humans observe objects. Google [24] proposed a novel recurrent neural network model, which extracted information from images or videos by adaptively selecting regions or position sequences and only processing the selected areas with high resolution. Google [25] also presented an attention-based model for identifying multiple objects in an image. In addition to the research on the attention mechanism algorithm, many scholars apply the attention mechanism to image classification. Haut et al. [26] proposed a new visual attention-based classification algorithm. Yang [27] proposed a RetinaNet model based on attention mechanism to match and classify the target ship accurately. In our previous model for navigation mark image classification [12], an attention mechanism based on three scale fusion of feature map was proposed to locate the area of attention and obtain characteristic.

However, the attention mechanism weakens the contour features. The results of the previous study [12] show that the RMA model has misclassification due to inconsistent appearance. The contour accentuation method can correct these problems [28]. In some fields, this method was widely used. Shotton [29] proposed a new type of automatic visual recognition system based on local contour features, which can locate objects in space and scale. It also confirmed that contour was a powerful hint for the multiscale and the multitype visual object recognition. Lin [30] also developed a new technology for detecting fruits in natural environments based on contour information. Their experiments showed that the proposed method was competitive for most types of fruits in natural environments, such as green, orange, circular, and nonround. To obtain higher accuracy of ship recognition, a contour accentuation method combined a ship recognition method based on transfer learning was proposed to analyse the ship images to detect the ship types. The actual results showed that the contour accentuation method with the transfer learning could obtain higher accuracy in ship image recognition [31]. Obviously, contour features are helpful to visual recognition. Therefore, in this paper, contour accentuation was expected to complement the affect of attention mechanism, and it was combined into the RMA model for navigation mark classification to further improve accuracy.

Recently, there are some intelligent information systems were reported about navigation mark management and service [32–34]. However, these systems were mainly developed based on telemetry and remote control; their identification mechanism of navigation marks is different from image recognition. Qi et al. [35] proposed a maritime navigation mark system based on electromagnetic waves, and Zhang [36] proposed a navigation mark communication system based on WLAN. These systems mainly provide information service of navigation marks by position instead of visual recognition. In this paper, a novel intelligent service system for image recognition of navigation marks was developed, and it orients to the application scenarios from camera.

3. Classification Models for Navigation Marks

This section firstly introduces the classification model of navigation marks called ResNet-Multiscale-Attention (RMA) model, then describes how to combine the RMA model with contour accentuation.

3.1. The ResNet-Multiscale-Attention (RMA) Model. In the daytime, navigation marks can be recognized by their shape,

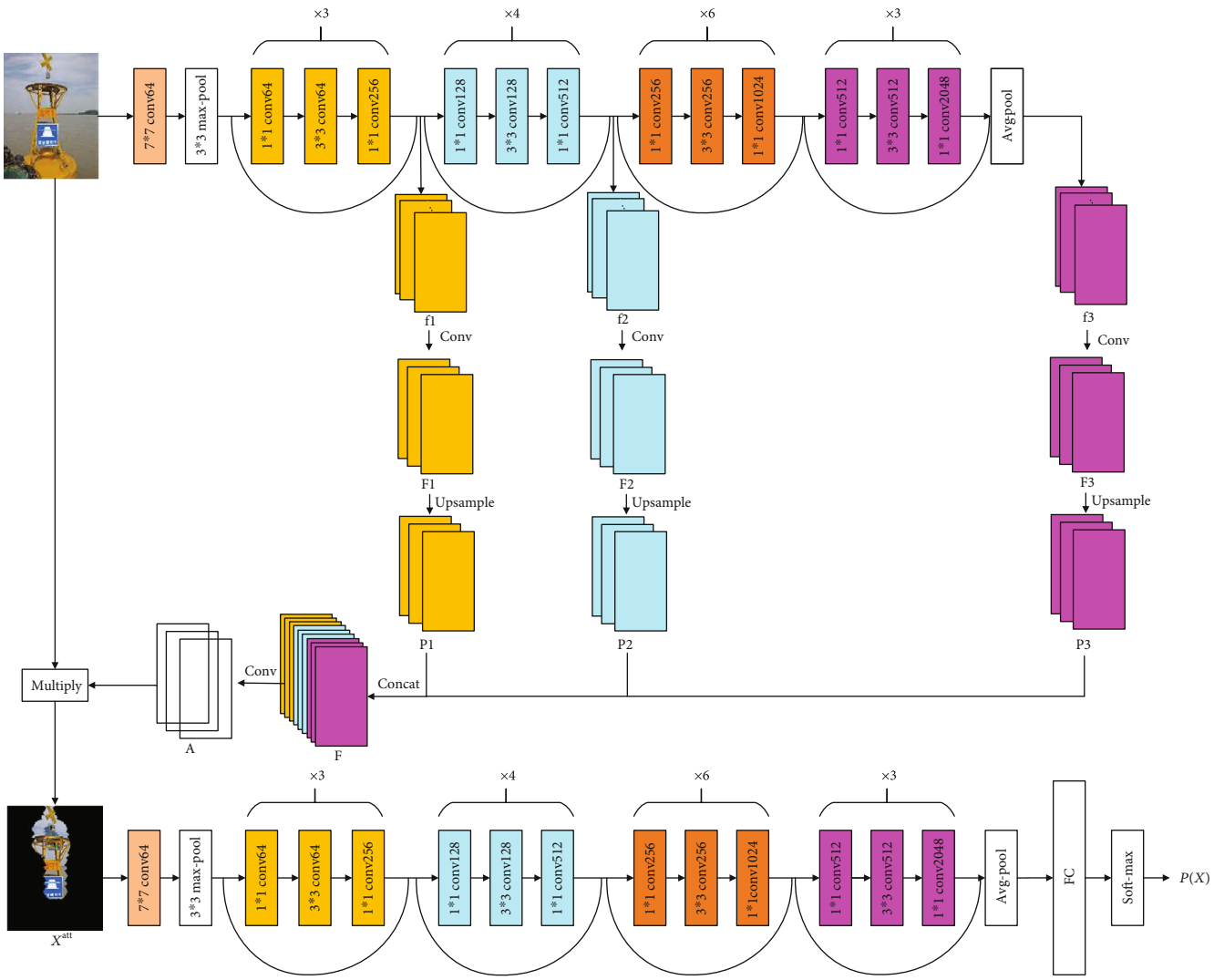
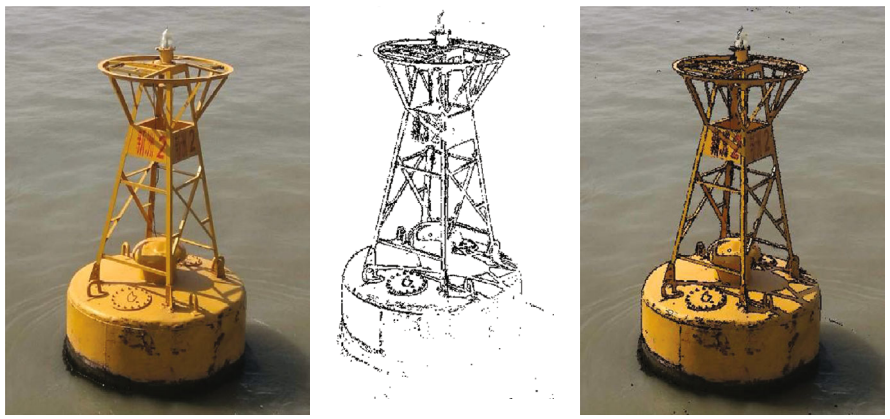


Figure 1: Network structure of the RMA [12].



(a) (b) (c)

Figure 2: Contour accentuation of navigation mark image.

color, and other auxiliary features. However, some kinds of navigation marks have a similar contour with subtle differences. Accordingly, for the visual navigation mark image recognition, the fine-grained image classification method is better than the general-level ones. Generally, in the deep neural networks of classification, low-level features have less

semantic information but more information about the target's position. Instead, high-level features have more semantic information but less detailed information about the target's position. General-level models usually do not perform well in fine-level tasks with the high-level features [12].

To tackle the fine-grained classification of navigation marks, a model called RMA was proposed in which the ResNet-50 was enhanced by adding a multiple scale attention mechanism [12]. As shown in Figure 1, in the network structure of RMA, the images of navigation mark were enhanced firstly by an improved ResNet-50, then classified by the second ResNet-50. The first ResNet-50 layer is designed as an attention matrix to capture the attention regions. Three-channel feature maps (f_1, f_2, f_3) from different stages of ResNet-50 represent three detail scales; there are integrated to form an attention matrix F by Convolution, Upsample, and Concat processes. And the F is then multiplied with element-wise of the input image to highlight the favourable classification area. The second ResNet-50 layer performs classification task and outputs final probabilities $p(x)$ of all navigation mark types.

Experiment results on a navigation mark image dataset showed that the RMA had classification accuracy about 95.98%, which was better than 94.14% of the ResNet-50.

3.2. RMA Model with Contour Accentuation. Contour features are helpful to visual recognition by enhancing target in the image, which was verified in many types of research and our other experiment about ship recognition. The multiple scale attention mechanism of RMA is aimed at locating the target's region, and the objective of contour is the enhancement of the target's features. In this paper, the contour accentuation method is considered to be combined into the RMA model for further improving the classification accuracy of navigation marks.

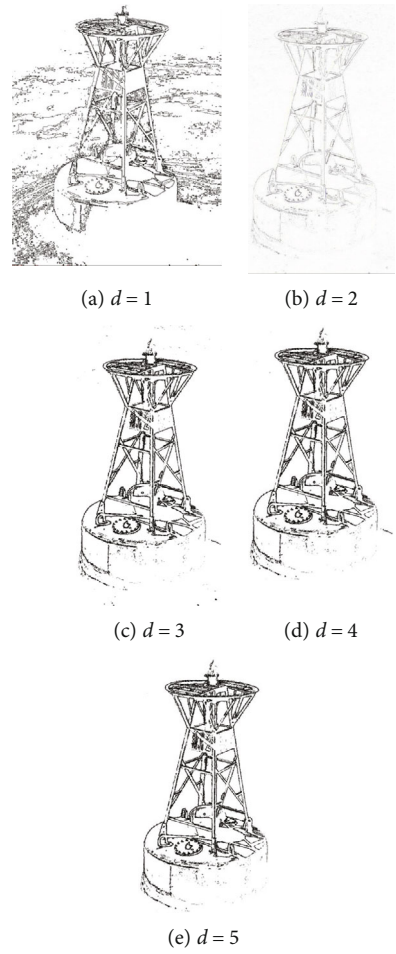
The original image of navigation mark as shown in Figure 2(a) is a $m \times n$size of color image, with red (R), green (G), and blue (B) three color channels. Its pixel matrix can be denoted as $\mathbf{C} = \{c_{1,1,1}, c_{1,1,2}, c_{1,1,3} \cdots, c_{x,y,1}, c_{x,y,2}, c_{x,y,3} \cdots, c_{m,n,1}, c_{m,n,2}, c_{m,n,3}\}$, where the value of red, blue, and green color in the pixel position (i, j) are $c_{i,j,1}$, $c_{i,j,2}$, and $c_{i,j,3}$, respectively.

$$D(P_{i,j}, P_{x,y}) = \prod_{1}^{3} \left| C_{i,j,k} - C_{x,y,k} \right|. \tag{1}$$

The function $D(P_{i,j}, P_{x,y})$ in Equation (1) is defined to measure the color difference between pixel (i, j) and pixel (x, y). If the color difference with all its neighbours is more significant than a critical value d, the pixel can be regarded as a contour point. Otherwise, it is not on contour. So, by Equation (2), the pixels on contour are set to black, otherwise set to white. The contour of navigation mark image can be captured in Figure 2(b).

$$f(P_{i,j}, P_{x,y}) = \begin{cases} 0, \text{if } \prod_{x=i-1}^{i+1} \prod_{y=j-1}^{j+1} D(P_{i,j}, P_{x,y}) \geq d, \\ 1, \text{otherwise.} \end{cases} \tag{2}$$

Furthermore, by Equation (3), which keeps the original color of pixels that is not on contour instead of white, an image with contour accentuation $\mathbf{C}' = \{c'_{1,1,1}, c'_{1,1,2}, c'_{1,1,3} \cdots, c'_{x,y,1}, c'_{x,y,2}, c'_{x,y,3} \cdots, c'_{m,n,1}, c'_{m,n,2}, c'_{m,n,3}\}$ can be obtained. In Figure 2(c), the navigation mark is enhanced by the contour features obviously.



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$ (e) $d = 5$

Figure 3: Contour accentuation results.

$$C'_{i,j,k} = \begin{cases} 0, \text{if } \prod_{x=i-1}^{i+1} \prod_{y=j-1}^{j+1} D(P_{i,j}, P_{x,y}) \geq d, \\ C_{i,j,k}, \text{otherwise.} \end{cases} \tag{3}$$

To combine the contour accentuation method with the RMA model, the contour accentuation algorithm can be used as an image preprocessing method, and the RMA model adopts the navigation images with contour accentuation as inputs directly.

4. Experiments and Results

To validate the effectiveness of the RMA model with contour accentuation, a navigation mark image dataset is firstly preprocessed with contour accentuation and then trained and tested with the RMA model.

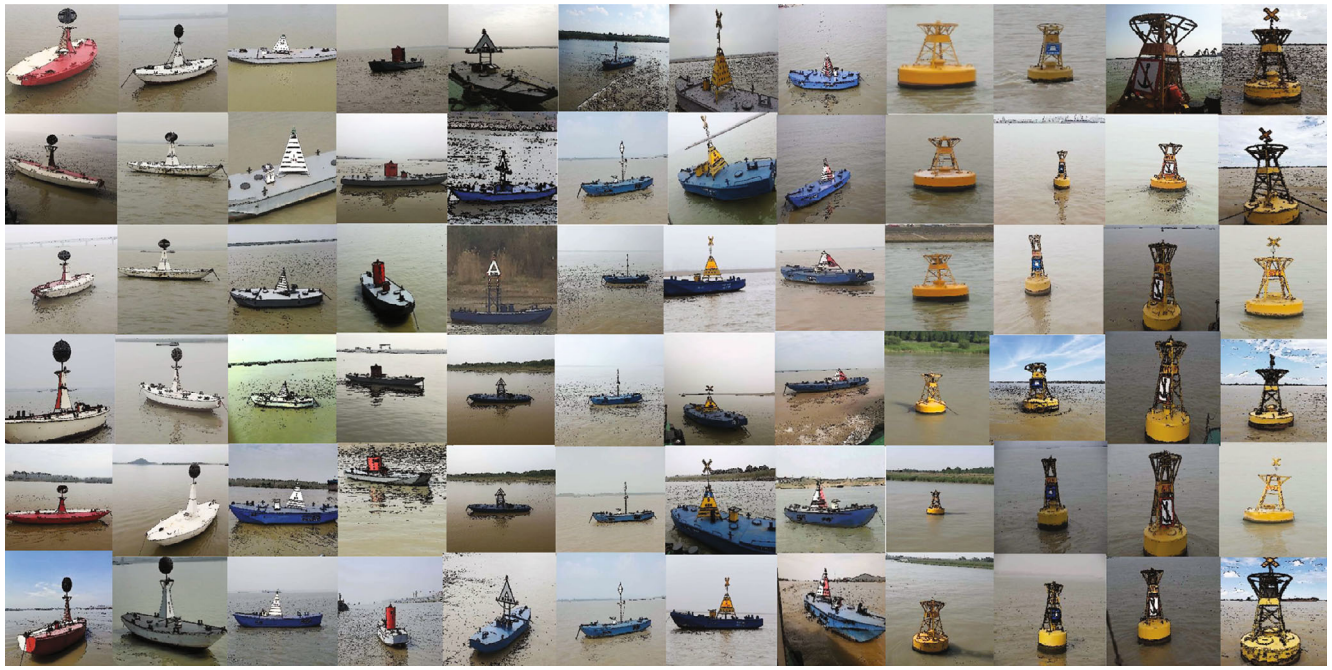


Figure 4: Dataset of navigation mark images with contour accentuation.

4.1. Dataset. A total of 10260 images of 42 kinds of navigation marks in the Yangtze River are collected. All images are clipped into a uniform size of 240 ∗ 240 to form an original dataset, and then, they are preprocessed with contour accentuation to create a contour enhanced dataset.

In Equation (2), critical value d is an important factor which determine the extraction effect of navigation mark's contour. After contour accentuation, not only the contour of navigation mark is enhanced, other features in the background such as the wave, mark's shadow, and reflection on the water surface also may be outlined in some extent, as Figure 3(a) showed, which will act as noises to disturb the recognition task. So, in order to eliminate the interference noises in the background and highlight the navigation mark to the maximum, the d should be chosen carefully.

From 1 to 5, the affection of different d was investigated. As Figure 3 showed, when d is less than 3, the noises are obvious (Figures 3(a) and 3(b)), and when d equals 3, the contour of navigation mark become sharp and clear, and the noises in background are suppressed (Figure 3(c)), when d is great than 3, the contour almost keep unchanged (Figures 3(d) and 3(e)).

Therefore, finally chose $d = 3$, and all images in original dataset were performed contour accentuation to form a new dataset. Figure 4 shows part images of the new contour enhanced dataset, in which "DCZYTHFB," "ZADCCMFB," "ZAZVXCMFB," "YAGXCMFB," "SDGGXFB," etc. are the labels of different kinds of navigation marks. And, in both original and contour enhanced datasets, each type images are divided into training and testing parts according to the ratio of 8 : 2.

4.2. Training Details. The RMA model is implemented by Python 3.7, the deep learning framework of TensorFlow 2.0, and trained in a workstation with two graphics cards of NVIDIA GeForce GTX 1080.

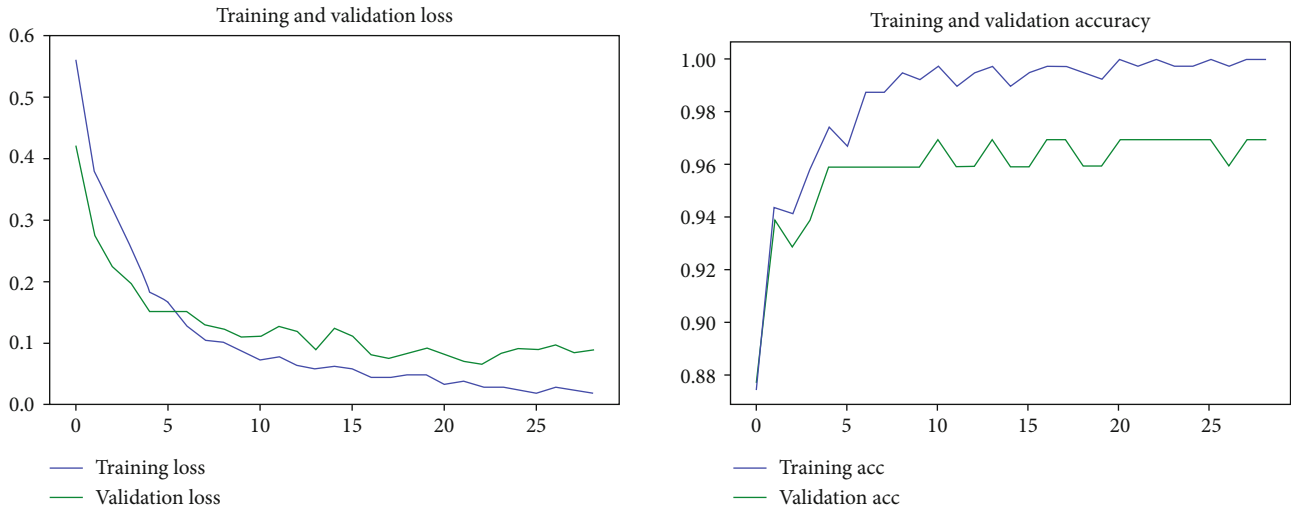
Since the number of images of the different navigation types in the dataset is unbalanced, the loss function is designed as Equation (4).

$$\text{loss} = -\sum_{i} w \times y_i \times \log\left(\text{logits}_i\right) + \left(1 - y_i\right) \times \log\left(1 - \text{logits}_i\right). \tag{4}$$

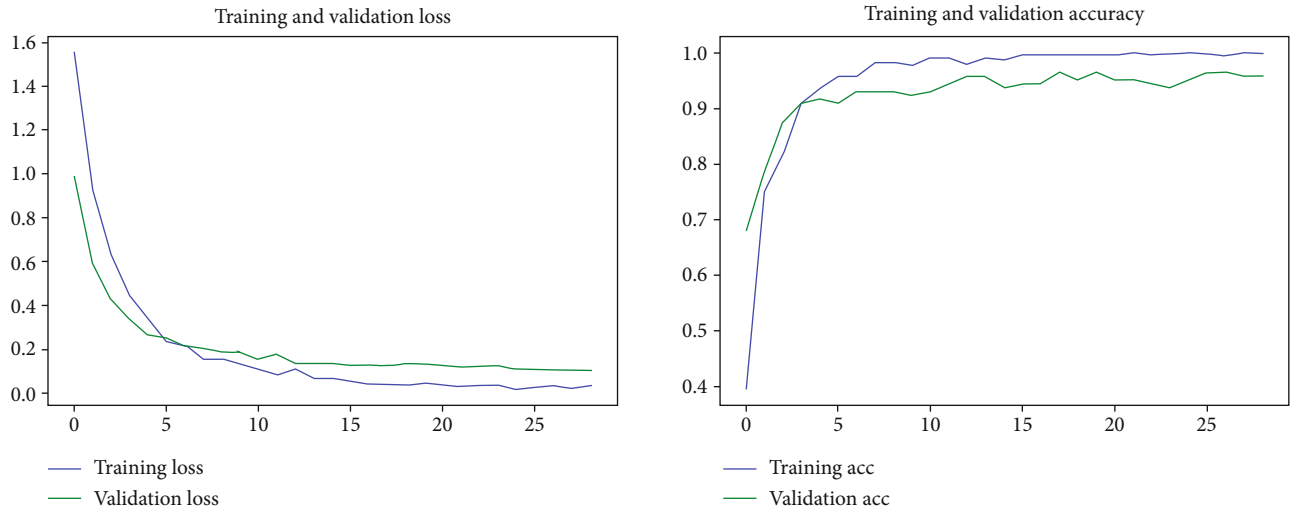
The parameter w is a calculation factor in advance based on the datasets, which has a more significant contribution to the loss function of the fewer types of samples. For the purpose of making the model converge faster, an SGD optimizer with momentum was used. For comparison, experiments were carried out on both the original dataset and the contour enhanced dataset.

Figures 5(a) and 5(b) show the loss and accuracy curves of the RMA model on the two datasets, respectively. Furthermore, to verify the effect of contour accentuation, the classification accuracy of six different deep learning network structures on the two datasets were also investigated.

4.3. Experimental Results. Table 1 shows all the experimental results. It can be found all models have slightly higher accuracy on the contour enhanced dataset than on the original dataset, and RMA has higher accuracy than ResNet-50 and other models on both two datasets. The results verified that contour accentuation could improve the classification accuracy generally. Moreover, the results also indicated that, for RMA, contour accentuation and multiple scale attention mechanism could complement each other well and improve the accuracy further.



(a) The loss and accuracy curves of the RMA model on navigation mark images



(b) The loss and accuracy curves of the RMA model on navigation mark images with contour accentuation

Figure 5: Loss curves and accuracy curves.

Table 1: The comparison of accuracy among different models.

Model	Original dataset	Contour enhanced dataset
GoogleNet	0.8881	0.8954
VGG-16	0.8973	0.9017
VGG-19	0.9035	0.9073
AlexNet	0.9091	0.9101
SqueezeNet	0.9208	0.9252
ResNet-50	0.9414	0.9466
RMA	0.9598	0.9653

The confusion matrices for misclassified images are shown in Figure 6. The red number indicates the number of errors in image classification. Rows are predicting types, and column types are reals. The matrices include 678 images of 12 classes in the test dataset. The results show that the misclassification is mainly caused by the subtle differences between classes such as "DCZYTHFB" and "ZADCCMFB," "GXJXFB" and "SDGXCMFB," "GXJXFB" and "ZADCCMFB," "SDGXGXFB" and "ZAZVXCMFB," and "ZADCCMFB" and "DCZYTHFB." The comparison results of Figures 6(a) and 6(b) show that the RMA model with contour accentuation reduces the total number of misclassified images from 14 to 7, with reduction in most types. The results show that contour accentuated model is more significant for classification results of navigation marks with different contours such as "GXJXFB" and "ZVXZYTHFB," "SDGXGXFB" and "ZUXZYFB," "SDGXGXFB" and "ZVXZYTHFB," "ZADCCMFB" and "ZUXTSFB," "ZUXZYFB" and "ZUXWXSYFB," and "ZVXZYTHFB" and "ZVXWXSYFB."

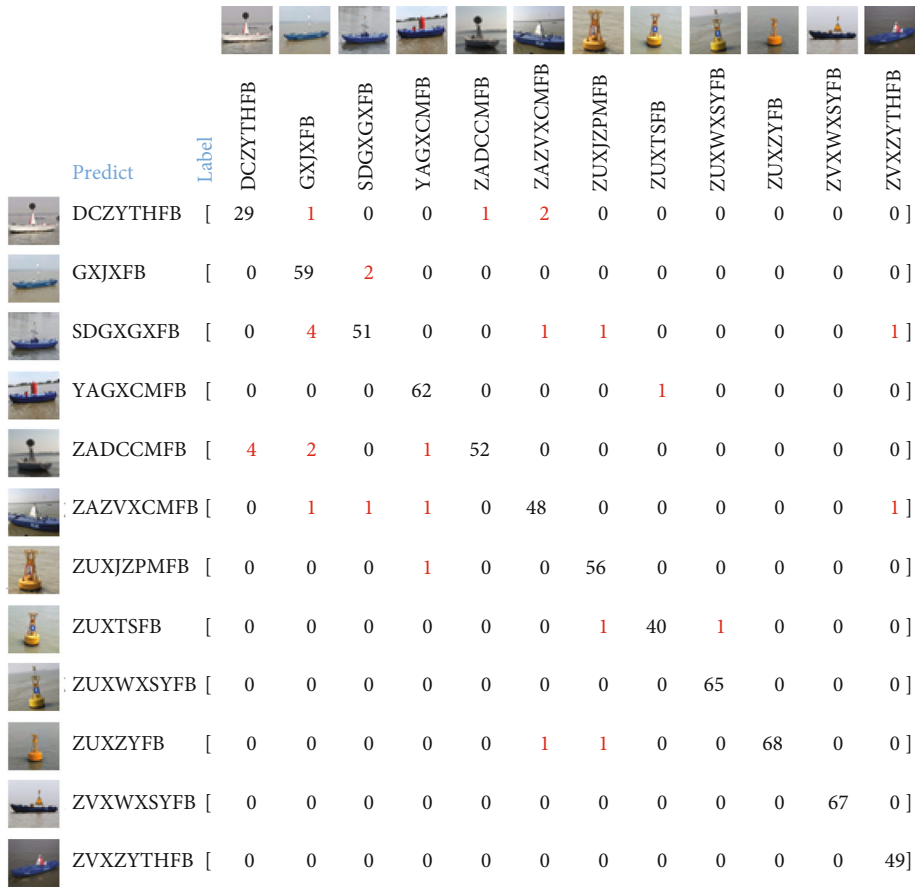
In order to verify the effect of the contour accentuation mechanism, the abovementioned navigation mark types with more improved accuracy were investigated further. As shown in Figure 7, the extracted contours of these navigation marks are clear, and there are few environment noises; the enhanced features by contour will help the RMA model to pay more

Predict \ Label	DCZYTHFB	GXJXFB	SDGXGXFB	YAGXCMFB	ZADCCMFB	ZAZVXCMFB	ZUXJZPMFB	ZUXTSFB	ZUXWXSYFB	ZUXZYFB	ZVXWXSYFB	ZVXZYTHFB
DCZYTHFB	30	0	0	0	3	0	0	0	0	0	0	0
GXJXFB	0	60	1	0	0	0	0	0	0	0	0	0
SDGXGXFB	0	4	52	0	0	1	1	0	0	0	0	0
YAGXCMFB	0	0	0	62	0	0	0	1	0	0	0	0
ZADCCMFB	4	3	0	0	50	0	0	0	0	0	0	0
ZAZVXCMFB	0	0	3	1	0	48	0	0	0	0	0	0
ZUXJZPMFB	0	0	0	1	0	0	56	0	0	0	0	0
ZUXTSFB	0	0	0	1	1	0	1	38	0	1	0	0
ZUXWXSYFB	0	0	0	0	0	0	0	0	64	1	0	0
ZUXZYFB	0	0	1	0	0	0	0	1	0	68	0	0
ZVXWXSYFB	0	0	0	0	0	0	0	0	0	0	66	1
ZVXZYTHFB	0	1	1	0	0	0	0	0	0	0	0	47

Original dataset

(a) Original dataset

Figure 6: Continued.



(b) Contour enhanced dataset

Figure 6: Misclassified image confusion matrix.

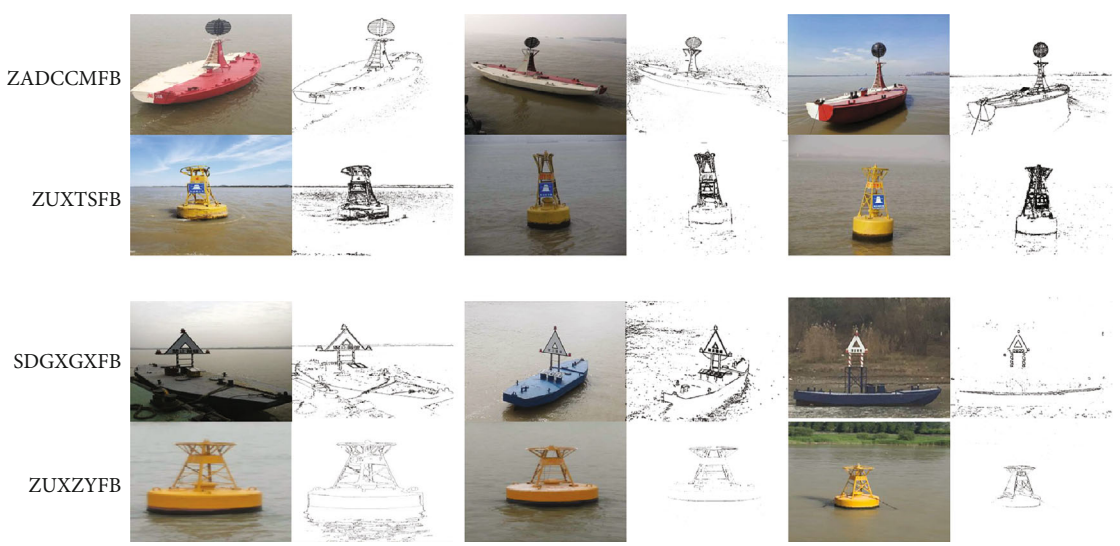


Figure 7: Visualization of the effect of contour accentuation.

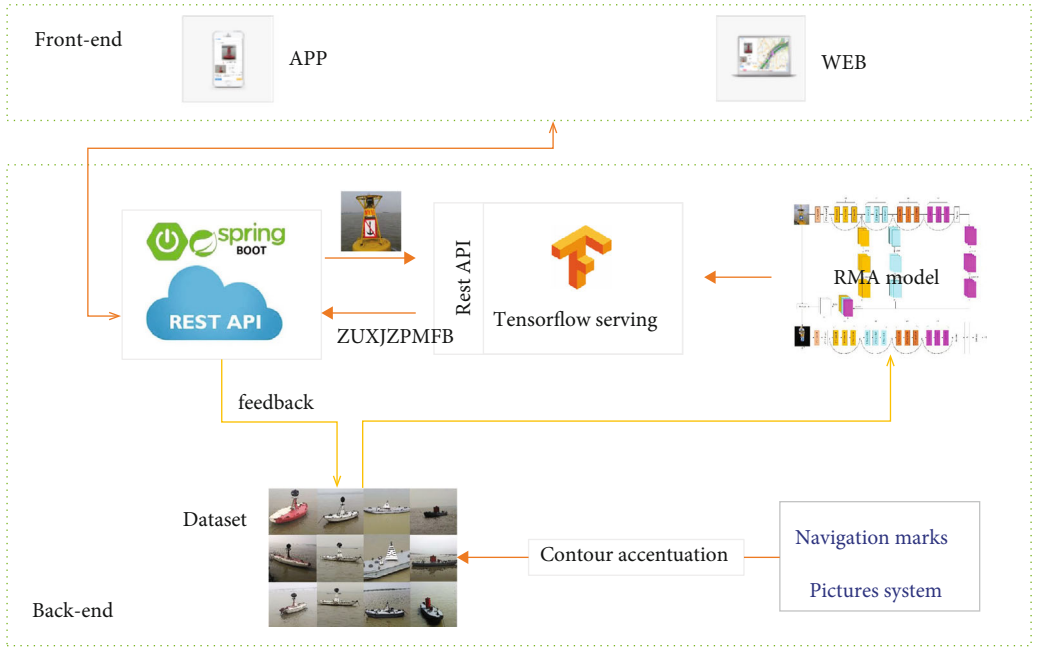


Figure 8: System architecture.

attention on navigation mark; furthermore, they will enlarge the distinguish between different types. This explained why the contour accentuation can improve the classification accuracy of navigation marks.

5. Application of Intelligent Recognition of Navigation Marks

Based on the RMA model with contour accentuation, an intelligent recognition system of navigation marks was developed for Changjiang Nanjing Waterway Bureau. The system has an architecture of front-end and back-end separation shown in Figure 8; the front-end focuses on client page (WEB or APP) rendering. In contrast, the back-end focuses on business logic, and they interact through the interface (REST APIs).

In the back-end, there are three platforms which are deployed independently but interacted with each other through an interface. The web service platform is developed and deployed based on the framework of Spring Boot. It interacts with front-end directly, accepts and transforms the image of request into required size and format, then sends it to the recognition module and gets recognition result. In the recognition module, TensorFlow Serving is used to deploy the RMA models, and a REST API for navigation mark recognition based on the latest model is exposed to the web service platform. In the training module, the RMA will be trained periodically in TensorFlow. Simultaneously, the dataset was enlarged by the image collection process of the digital waterway system (the production system for channel maintenance in Nanjing Waterway Bureau), and the model will be saved with a version number, updated, and loaded into TensorFlow Serving.

The front-end can be a variety of clients, Web, APP, or WeChat Mini Program. The clients accept the uploaded image, send it to the web service platform, and get the responses of recognition and rendering them on the page as Figure 9 showed.

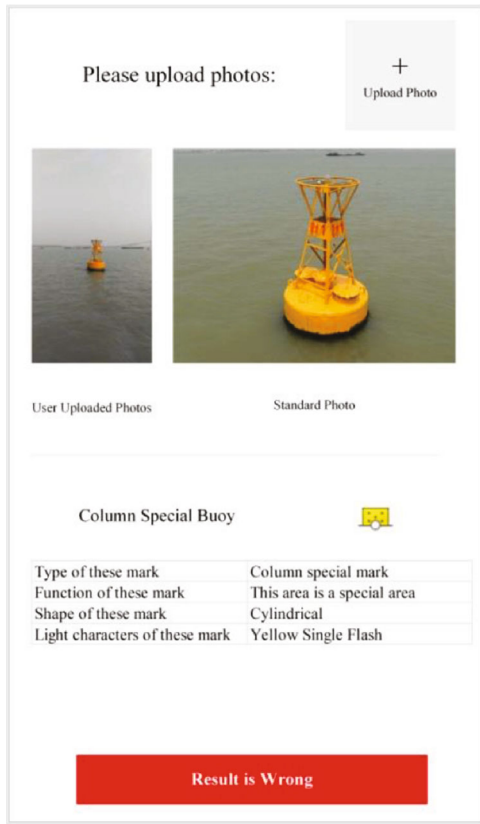


Figure 9: Navigation mark recognition page.

6. Conclusions and Future Work

This paper applies deep learning technology to study the navigation mark image recognition. It proposes a navigation mark classification model based on the combination of multiscale attention mechanism and contour accentuation. The effect of multiple scale attention mechanisms for improving classification accuracy has been validated in our previous works about the RMA model. This paper mainly focused on the impact of contour accentuation. Experimental results on 10260 navigation mark images showed that by enhancing the contour of the object, contour accentuation could improve the image classification accuracy of most general classification models. It also improves the RMA model well and increases the classification accuracy from 95.98% to 96.53%.

Based on the improved classification model, this paper further developed an intelligent service system for the recognition of navigation marks. The system has a flexible architecture based on front-end and back-end separation. It is connected with the digital waterway system to obtain a continuously updated dataset and then realized an automatic navigation mark recognition service including dataset preparation, model training, model deployment, and model update.

In the future, the value of d in the proposed contour accentuation algorithm could be optimized for different light conditions. In addition, in order to further enhance the accuracy of navigation mark image classification, the adversarial neural network can be studied and applied to the fine-grained classification of navigation mark images.

Data Availability

Access to the image dataset of navigation marks used to support the findings of this study is restricted, because it belongs to a third party, the Changjiang Nanjing Waterway Bureau of the People's Republic of China.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the Fundamental Research Funds for the Central Universities under Grant 3132019400. Thanks are due to the Changjiang Nanjing Waterway Bureau of the People's Republic of China for providing the image dataset of navigation marks and application scenario of the research results. This work was also partially supported by the National Natural Science Foundation of China (Nos. 61906043, 61902313, 61902072, 62002063, 61877010, 11501114, and 11901100), the Fujian Natural Science Funds (Nos. 2020J05112, 2020J05111, 2020J01498, and 2019J01243), the Funds of Education Department of Fujian Province (No. JAT190026), and the Fuzhou University (Nos. 0330/50016703, 0330/50009113, 510930/GXRC-20060, 510872/GXRC-20016, 510930/XRC-20060, 510730/XRC-18075, 510809/GXRC -19037, 510649/XRC-18049, and 510650/XRC-18050).

References

[1] I. Im, D. Shin, and J. Jeong, "Components for smart autonomous ship architecture based on intelligent information technology," *Procedia Computer Science*, vol. 134, pp. 91–98, 2018.

[2] *International Dictionary of Marine Aids to Navigation*October 2020, https://www.ialaaism.org/wiki/dictionary/index.php/Aid_to_Navigation.

[3] X. Guo, "Application and management of AIS aids to navigation," *Ship Electronic Engineering*, vol. 36, no. 6, pp. 54–58, 2016.

[4] J. Pandy and K. Hasegawa, "Autonomous navigation of catamaran surface vessel," in *2017 IEEE Underwater Technology (UT)*, pp. 1–6, Busan, South Korea, 2017.

[5] J. Zhuang, L. Zhang, S. Zhao, J. Cao, B. Wang, and H. Sun, "Radar-based collision avoidance for unmanned surface vehicles," *China Ocean Engineering*, vol. 30, no. 6, pp. 867–883, 2016.

[6] A. Garcia-Dominguez, "Mobile applications, cloud and big-data on ships and shore stations for increased safety on marine traffic; a smart ship project," in *2015 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1532–1537, Seville, Spain, 2015.

[7] Y. Tang and N. Shao, "Design and research of integrated information platform for smart ship," in *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pp. 37–41, Banff, AB, Canada, 2017.

[8] B. Liu, S. Wang, J. Zhao, and M. Li, "Ship tracking and recognition based on Darknet network and YOLOv3 algorithm," *Journal of Computer Applications*, vol. 39, no. 6, pp. 1663–1668, 2019.

[9] H. Fu, Y. Li, Y. Wang, and P. Li, "Maritime ship targets recognition with deep learning," in *2018 37th Chinese Control Conference (CCC)*, pp. 9297–9302, Wuhan, China, 2018.

[10] Z. Li, L. Zhao, X. Han, M. Pan, and F. J. Hwang, "Lightweight ship detection methods based on YOLOv3 and DenseNet," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4813183, 10 pages, 2020.

[11] International Maritime Organization, "International Convention on Standards of Training, Certification and Watchkeeping for Seafarers, 1978, as amended in 1995," 1997.

[12] M. Pan, Y. Liu, J. Cao, Y. Li, C. Li, and C. Chen, "Visual recognition based on deep learning for navigation mark classification," *IEEE Access*, vol. 8, pp. 32767–32775, 2020.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–14, San Diego, CA, USA, 2015.

[15] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[17] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, HI, USA, 2017.

[18] X. Yao, X. Wang, S. Wang, and Y. Zhang, "A comprehensive survey on convolutional neural network in medical image analysis," *Multimedia Tools and Applications*, 2020.

[19] M. Wang and W. Deng, "Deep face recognition: a survey," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 471–478, Parana, Brazil, 2018.

[20] Q. Shi, W. Li, F. Zhang, W. Hu, X. Sun, and L. Gao, "Deep CNN with multi-scale rotation invariance features for ship classification," *IEEE Access*, vol. 6, pp. 38656–38668, 2018.

[21] Q. Oliveau and H. Sahbi, "From transductive to inductive semi-supervised attributes for ship category recognition," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4827–4830, Valencia, 2018.

[22] H. C. Shin and K.-I. Lee, "Classification maritime vessel image utilizing a region of interest extracted and convolution neural network," *Journal of Korean Institute of Intelligent Systems*, vol. 29, no. 4, pp. 321–326, 2019.

[23] B. Solmaz, E. Gundogdu, V. Yucesoy, A. Koç, and A. A. Alatan, "Fine-grained recognition of maritime vessels and land vehicles by deep feature embedding," *IET Computer Vision*, vol. 12, no. 8, pp. 1121–1132, 2018.

[24] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2204–2212, 2017.

[25] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *International Conference on Learning Representations*, pp. 1–10, San Diego, CA, USA, 2015.

[26] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 8065–8080, 2019.

[27] T. Yang, Z. Chen, Y. Lv, Y. Wu, and B. Hua, "Multi-resolution ocean target detection method based on deep learning," *Electronics Optics & Control*, pp. 1–7, 2020, http://kns.cnki.net/kcms/detail/41.1227.TN.20200817.1258.020.html.

[28] J. Victorino and F. Gómez, "Contour analysis for interpretable leaf shape category discovery," *Plant Methods*, vol. 15, no. 1, 2019.

[29] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1270–1281, 2008.

[30] G. Lin, Y. Tang, X. Zou, J. Cheng, and J. Xiong, "Fruit detection in natural environment using partial shape matching and probabilistic Hough transform," *Precision Agriculture*, vol. 21, no. 1, pp. 160–177, 2020.

[31] C. Chen, Y. Zhang, W. Guo, M. Pan, L. Lyu, and C. Lin, "Contour accentuation for transfer learning-based ship recognition method," in *Proceedings of the Web Conference 2020 (WWW'20)*, New York, NY, USA, 2020.

[32] S. Beatriz, C. Nicoleta, and F. Francisco, "Artifical intelligence to determine if liquified natural gas in short sea shipping is a social bet," *Ingeniería y Desarrollo*, vol. 36, pp. 418–436, 2018.

[33] D. S. Cristea, L. M. Moga, M. Neculita, O. Prentkovskis, K. M. D. Nor, and A. Mardani, "Operational shipping intelligence through distributed cloud computing," *Journal of Business Economics and Management*, vol. 18, no. 4, pp. 695–725, 2017.

[34] M. Sun, "Research on management informationization of inland waterway," *People's Transportation*, vol. 5, p. 76, 2019.

[35] S. Qi, H. Zhang, and J. Tian, "Research on wireless location method of short baseline marine beacons based on phase measurement," in *2018 2nd IEEE Advanced Information Management,Communicates,Electronic and Automation Control Conference (IMCEC)*, pp. 1123–1129, Xi'an, 2018.

[36] J. Zhang, "Research of application of communication technology of WLAN based on ship," in *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, Deng Feng, China, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6655125, 13 pages
https://doi.org/10.1155/2021/6655125

Research Article

An Improved Unsupervised Single-Channel Speech Separation Algorithm for Processing Speech Sensor Signals

Dazhi Jiang,[1] **Zhihui He**,[1] **Yingqing Lin**,[1] **Yifei Chen**,[1] **and Linyan Xu**[2]

[1]*Department of Computer Science, Shantou University, China 515063*
[2]*Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy 20156*

Correspondence should be addressed to Dazhi Jiang; dzjiang@stu.edu.cn

Received 29 December 2020; Revised 24 January 2021; Accepted 1 February 2021; Published 27 February 2021

Academic Editor: Xingsi Xue

As network supporting devices and sensors in the Internet of Things are leaping forward, countless real-world data will be generated for human intelligent applications. Speech sensor networks, an important part of the Internet of Things, have numerous application needs. Indeed, the sensor data can further help intelligent applications to provide higher quality services, whereas this data may involve considerable noise data. Accordingly, speech signal processing method should be urgently implemented to acquire low-noise and effective speech data. Blind source separation and enhancement technique refer to one of the representative methods. However, in the unsupervised complex environment, in the only presence of a single-channel signal, many technical challenges are imposed on achieving single-channel and multiperson mixed speech separation. For this reason, this study develops an unsupervised speech separation method CNMF+JADE, i.e., a hybrid method combined with Convolutional Non-Negative Matrix Factorization and Joint Approximative Diagonalization of Eigenmatrix. Moreover, an adaptive wavelet transform-based speech enhancement technique is proposed, capable of adaptively and effectively enhancing the separated speech signal. The proposed method is aimed at yielding a general and efficient speech processing algorithm for the data acquired by speech sensors. As revealed from the experimental results, in the TIMIT speech sources, the proposed method can effectively extract the target speaker from the mixed speech with a tiny training sample. The algorithm is highly general and robust, capable of technically supporting the processing of speech signal acquired by most speech sensors.

1. Introduction

As information technology is advancing and 5G technology is being popularized, Internet of Things (IoT) devices and sensors will be increasingly created, which will undoubtedly change the way human beings live. Moreover, sensor networks are being progressively studied [1–3]. It is predicted that in the next decade, billions of IoT and sensor devices will generate massive data for applications in smart grid, smart home, electronic health, industry 4.0, etc. It is foreseeable that intelligent speech systems will be critical to the mentioned areas. With the rapid growth of data volume, large-scale problems should be urgently solved effectively [4, 5], while more opportunities are brought. Speech sensor networks, an important part of IoT, will have many application needs. However, in real-world scenarios, the data acquired by speech sensors are often disturbed by noise. Thus, low-noise and effective speech data should be urgently obtained.

With the increasing number of speech sensors, reliable speech separation technology is required [6–8]. High reliable speech separation technology is capable of achieving effective speech recognition, so the needs of human hearing can be satisfied. Speech separation originates from blind source separation (BSS) [9]. The core goal of this technology is to separate the source signal from the measured mixed signal. In the blind source analysis task, the target speech should be separated from the mixed speech in a single channel, which is very difficult to achieve. Single-channel speech separation is a hotspot in the current research. Many algorithms are proposed for single-channel speech separation, but from the current research results, this problem is far from being well solved. We believe that the current challenges are mainly manifested as the following aspects.

(1) Strong noise and unknown number of sources still significantly enhance the performance of BBS. Indeed, most of the existing blind source separation algorithms have achieved ideal performance in high SNR (Signal-to-Noise Ratio) environment. In practical applications, the signal we collected may have been polluted by strong noise. Because of this, many reported algorithms in blind source separation are very likely to obtain poor separation performance and even cannot correctly deal with the severely distorted signal in extreme cases. For the mentioned reason, to obtain robust blind source separation algorithm, a more effective method is required to suppress the impact of noise. In addition, a more difficult problem is that the number of sources is unknown. On the whole, the number of sources should be assumed, whereas in practical applications, information on the number of sources is not available, which cannot be ignored [10]. Accordingly, blind estimation of the number of sources from the received mixed signal cannot effectively obtain the ideal BSS performance

(2) The processing complexity of single-channel speech separation is higher than that of multi-input speech separation. In numerous practical applications, the challenge of blind source separation is that only one sensor is available, namely, SCBSS (single-channel and blind source separation) [11–13]. It uses only a single receiver sensor to receive the observed signal and then uses the signal to recover each source signal. Generative adversarial network (GAN) is an excellent representative of deep learning algorithms and is also used in SCBSS due to its advantages in fitting data distribution (e.g., 1D speech signal separation [14–16]). However, the performance of GAN is limited by the unknown number of source signals, complex forms of dialogue, serious noise pollution, and difficulty in obtaining prior information in advance. Such a type of SCBSS is characterized by unknown number of source signals, complex dialogue form, serious noise pollution, and difficulty in acquiring prior information in advance. To solve this type of problem, unsupervised learning method should be developed, whereas automatic analysis should be extremely difficult to realize based on unsupervised learning method (overall, single-channel speech only requires a single signal source, which is easier to achieve and more realistic than multichannel speech)

(3) The solution to solve BBS problem refers to employing supervised learning mechanism. The more representative is the deep learning method. It has been recently found that deep learning [17, 18] has achieved remarkable success in many speech processing fields with its excellent learning performance. The representative technology is DNN-HMM hybrid structure [19, 20], replacing the conventional acoustic modeling based on GMM and HMM. In single-channel speech separation, a method based on DNNs [21, 22] has been proposed to separate the target speaker from the mixed speech. However, all deep learning algorithms use joint a decoding framework, which requires additional computational complexity. Moreover, deep learning algorithm needs considerable training data, which is difficult to extend to small data sets and unsupervised speech separation scenarios

To reduce the above challenges, an unsupervised speech separation method CNMF+JADE is proposed in this study, i.e., a hybrid method combined with Convolutional Non-Negative Matrix Factorization [23, 24] and Joint Approximative Diagonalization of Eigenmatrix [25]. This study is aimed at performing efficient processing for the highly noisy signal data acquired by the speech sensor to achieve better separation performance. CNMF refers to a nonnegative matrix decomposition method proposed for speech signal processing. The method adopts a 2D time-frequency basis instead of the 1D basis vector in the original nonnegative matrix decomposition, while it ensures the decomposition result to be nonnegative matrix decomposition. Thus, it effectively carries the correlation between local frames of speech signals [26]. JADE is recognized as an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrix, and it is an effective method for blind source separation. It exploits the feature that mutual accumulation is always zero when signals are independent and builds multiple fourth-order accumulation matrices for multivariate data. Lastly, the mentioned cumulant matrices are jointly diagonalized to solve for the final separated signals [27, 28]. For single-channel signal, CNMF+JADE can effectively separate the overlapped speech including the target speaker. Subsequently, CNMF+JADE with adaptive speech enhancement technology is adopted to further improve the speech quality of the target speaker. To solve the problem of SCBSS, the main innovations can be summarized below.

(1) In this study, CNMF and JADE are combined to solve the problem of single-channel speech separation. The algorithm is appropriate in extracting signals of interest from mixed signals. Specific to SNR (Signal-Noise Ratio), STOI (Short-Time Objective Intelligibility), and PESQ (Perceptual Evaluation of Speech Quality), the proposed CNMF+JADE, as compared with several speech separation methods (CNMF, CNMF +ICA), achieves satisfactory results, especially for single-channel mixed speech

(2) Given the scenario that the speech signal will get worse when the speech signal is enhanced after speech separation, an adaptive method is presented here based on wavelet transform to analyze the speech signal after CNMF+JADE separation, as an attempt to realize selective speech enhancement and increase the efficiency of speech enhancement

The rest of the study is organized as follows. In Section 2, some related studies on the study of single-channel speech separation are presented. In Section 3, the proposed algorithm is elucidated. In Section 4, a specific experimental

verification of the performance of the proposed algorithm is presented. Lastly, in Section 5, the conclusion and promising future research directions are drawn.

2. Related Work

As IoT technology is developing, intelligent voice system will have increasingly broad application prospects. In addition, single-channel blind speech separation (SCBSS) technology will arouse wide attention. At present, there are three main directions for SCBSS research:

(1) *Subspace Decomposition-Based Approach* [29]. Methods based on subspace decomposition primarily are aimed at identifying new descriptions. The mentioned new descriptions can often effectively extract perceptive meaningful component sources from complex mixtures [30]. Moreover, new descriptions can eliminate intrusions and reduce signal dimensionality, so redundant components can be avoided. The methods based on subspace decomposition are primarily well established in statistical and transformed data. For instance, in the literature [31], the effectiveness of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) methods in solving subspace decomposition problems has been verified. In fact, methods based on algebraic properties are more often used in dealing with subspace decomposition problems, including Non-negative Matrix Factorization (NMF) [32]. NMF is a classical time-frequency distribution method and is often used for single-channel speech separation [33–37]. Ref [38, 39] highlighted NMF as an unsupervised dictionary-based learning method that effectively helps solve various types of signal separation

(2) *Model-Based Approach*. In the first step of the model-based approach, each speaker in the model scene should be identified, and the gain in the blended frames should be determined. In fact, speaker recognition algorithms have been studied by many authors (e.g., Iroquois [40], Closed loop [41], and Adaptive Speaker Identification (SID) [42]). The next step is to choose an appropriate speech representation. The final step comprises the reconstruction of the speech signal frames, in which separated speech is produced. Overall, the reconstruction usually requires the construction of a hybrid estimator module that enables it to find a sufficient number of representative speech frames from the speaker model to rebuild a meaningful speech signal. However, mixture estimators are capable of significantly complicating the algorithm, so it is difficult to apply in real-time systems

(3) *Computational Auditory Scene Analysis- (CASA-) Based Approach*. CASA runs in two main stages, i.e., segmentation and grouping. The former comprises feature extraction, time-frequency analysis, and multitone tracking, while the latter includes the resynthesis of speech signals. To be specific, pitch tracking is an important technique when CASA is being used for SCBSS problem processing. Jin [43] and Tolonen [44] provided several pitch tracking methods that are used extensively. However, as impacted by the periodic nature of the grouping phase, it can only be limited to voiced speech segments. Moreover, the performance achieved by CASA-based methods tends to be affected by multi-pitch estimation for its dependence on pitch

Over the past few years, with the development of deep learning, researchers have suggested that the nonlinear processing and feature learning capabilities of deep models exhibit significant advantages in solving speech separation problems. For this reason, many models using deep learning for speech separation have been proposed (e.g., Deep Neural Network (DNN), deep stacking, Deep Stack Neural Network (DSN) [45], and other efficient deep learning models [46–50]). In addition, numerous deep learning algorithms have been proposed for single-channel speech separation [51–54]. The reason why deep learning is so effective in addressing with speech separation problems is that the speech separation problem is described as a supervised problem in the deep learning model. Thus, deep learning models can train and learn features from speech signals to effectively separate speech signals.

3. Methodology

In the present section, the methods we use for processing speech data are described, and a new algorithm with high generality and robustness is proposed, aiming to provide a general and efficient speech processing algorithm for the data acquired by speech sensors.

3.1. Speech Separation

(1) *CNMF*. Speech signals exhibit local interframe correlation and global interframe correlation. The conversion of local interframe correlation should consider two aspects, i.e., to ensure the continuity between frames of the converted voice channel spectrum, as well as to remove the source speaker features from the local interframe correlation and make it have the target speaker features. However, the conventional nonnegative matrix factorization does not consider the conversion of local frames. CNMF refers to a proposed nonnegative matrix decomposition method for speech signal processing. The method employs a 2D time-frequency basis instead of the 1D basis vector in the original nonnegative matrix decomposition while ensuring the nonnegativity of the decomposition result. Thus, the correlation between the local frames of the speech signal is carried effectively

The CNMF is expressed as follows:

$$Y \approx \sum_{t=0}^{T-1} A(t) \cdot \overrightarrow{X}^{t}, \tag{1}$$

where $Y \in M \times N$ and $X \in r \times N$ represent the time-frequency atoms and the corresponding time-varying gain coefficients, respectively. $(\cdot)^{i\rightarrow}$ denotes shifting the encoding matrix X by i units to the right in the form of column vectors and set the leftmost i column to 0.

In other words, the decomposition matrix Y is obtained by convolving a series of nonnegative fundamental matrices A and coefficient matrices X. The functions of CNMF are to find a series of fundamental matrices $A(t)$ and coefficient matrices X and then make the convolution result as close as possible to the target matrix Y.

In addition, the divergence $K - L$ acts as the cost function in CNMF:

$$D\left(Y|\widehat{Y}\right) = \sum_{i,j}\left(Y_{ij}\log\left(\frac{Y_{ij}}{\widehat{Y}_{ij}}\right) - Y_{ij} + \widehat{Y}_{ij}\right), \tag{2}$$

where $\widehat{Y}$ denotes the estimation of $\widehat{Y}$, and

$$\widehat{Y}_{ij} = \left(\sum_{t=0}^{T-1} A(t)\cdot \overset{t\rightarrow}{X}\right)_{ij}. \tag{3}$$

$K - L$ makes the maximum log-likelihood solution of solving the nonnegative matrices $A(t)$ and X under the Poisson noise assumption to describe the degree of approximation of $\widehat{Y}$ with respect to Y. The iterative function can be defined as follows.

$$X = X \otimes \frac{A(t)^T \cdot \left[\overset{\leftarrow t}{Y}/\widehat{Y}\right]}{A(t)^T \cdot E}, \tag{4}$$

$$A(t) = A(t) \otimes \frac{\left(Y/\widehat{Y}\right)\cdot \overset{t\rightarrow}{X^T}}{E \cdot \overset{t\rightarrow}{X^T}}, \tag{5}$$

where E indicating the matrix with all elements of 1 and $\otimes$ is the matrix element multiplication operator. When $T = 1$, i.e., $t = T - 1$ is 0, it will degenerate into the basic NMF decomposition. For each t, there is a basic matrix $A(t)$ corresponding to it.

3.1.1. JADE. The Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm is an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrices and an effective method for blind source separation. JADE mainly uses the diagonalization of Jacobi matrix to find the independent components, as an attempt to achieve the identification and separation of signals. Based on the characteristics of JADE mentioned above, JADE is introduced to effectively separate the acquired speech signals.

JADE algorithm first spheres the observed signal using an $n \times m$ spherization matrix to obtain the observation vector $u = [u_1, u_2, \cdots, u_N]^T$ for N channels. Then, let M be any $N \times N$ matrix, then the definition of the four-dimensional cumulant matrix $Q_u(M)$ of u is:

$$[Q_u(M)]_{ij} = \sum_{k=1}^{N}\sum_{l=1}^{N} K_{ijkl}(u)m_{kl}, i, j = 1, 2, \cdots, N, \tag{6}$$

where $K_{ijkl}(u)$ denotes the fourth-order cumulant of the i, j, k, and l components in the vector.

3.1.2. CNMF+JADE. However, in the same channel spectral matrix Y, the final $A(t)$ and X obtained by CNMF analysis are not the same when the initial values of $A(t)$ and X are different, i.e., the same time-frequency spectral matrix Y has multiple combinations of time-frequency bases and coding matrices. For the mentioned reason, if the parallel channel spectral matrices of the source and target speakers are analyzed independently by convolutional nonnegative matrix decomposition, the same encoding matrix that characterizes the content information is not ensured to be obtained. From the analysis described in Section 3.1.1, JADE is known as an adaptive batch independent component optimization algorithm based on multivariate fourth-order cumulative matrices and an effective method for blind source separation, capable of effectively identifying and separating signal, which achieves the obtained signals as identical as possible.

Accordingly, to efficiently process the speech signals collected by the speech sensors, a single-channel speech separation algorithm combining CNMF and JADE is proposed. The secondary separation process is performed on the speech signal separated by CNMF based on JADE. The role of CNMF+JADE algorithm is to separate the single-channel mixed speech and lastly acquire the separated speech signal of all speakers in the mixed speech. The algorithm exhibits strong generality and robustness, capable of technically supporting the processing of speech signals collected by most speech sensors. For instance, in the literature [55], several applications (e.g., beamforming, automatic camera steering, robotics, and surveillance) are processed with the speech separation method. In [56], a speech signal separation method is adopted for speech separation of noisy robust speech translation for general-purpose smart devices. It is foreseen that speech separation techniques are also critical to future applications of IoT technologies (e.g., driverless, smart home, and other applications involving sound conduction functions). For this reason, it is of great value and significance to propose more efficient speech separation algorithms (e.g., CNMF+JADE) as proposed in this study.

Lastly, the CNMF+JADE algorithm is described as follows.

The proposed algorithm is written in Algorithm 1, where $t_1, t_2, \cdots, t_N$ represent the set of all the pure speech signal data of the speaker waiting to be separated, $o_1, o_2, \cdots o_{N-1}$ denote the set of all the mixed speech employed as the training set, O is a mixed speech waiting to be separated, R_i is a random

Input: Speech signal dataset, $t_1, t_2, \cdots, t_N$, $o_1, o_2, \cdots o_{N-1}$ and O.
1: Initialize each parameter and variable:
T=$t_1, t_2, \cdots, t_N$, expresses the set of all the pure speech signal data of the speaker waiting to be separated,
H=$o_1, o_2, \cdots o_{N-1}$, expresses the set of all the mixed speech that is used as the training set,
O denotes a mixed speech waiting to be separated,
R_i is a random matrix.
2: **while** $i < N$ **do**
3: The speech data with the identical subscript $t_i$ando_io_i from the datasets T and H are selected to train CNMF.
4: The trained CNMF is employed to separate the mixed speech dataset O to determine $\widehat{s}_i$ and $\widehat{O}_i$.
5: The two speech signals acquired from 4 are mixed to obtain a two-channel speech signal and stored in R_i.
6: A secondary separation is conducted by adopting JADE to obtain $\widehat{s}_i$ and $\widehat{O}_i$ from R_i.
7: $\widehat{O}_i$ is used as the speech signal to be separated in the next round, and t_i and o_i are removed from the data sets T and H.
8: Obtain the final separated speech signal.
9: i = i +1.
10: **end while.**
Output: All of speaker's speech signals $s_1, s_2, \cdots, s_N$.

Algorithm 1: CNMF+JADE description.

matrix, and N represents the number of speech signals, i.e., the number of speakers. o_i denotes the corresponding speaker, as expressed in the dataset O.

$$o_i = O - \sum_{k=1}^{i} t_k, i = 1, 2, \cdots, N-1. \tag{7}$$

In fact, $o_1, o_2, \cdots o_{N-1}$ are very costly and difficult to obtain. Thus, in experiments, a speech signal different from the current target speaker is generally selected randomly from the dataset O to train CNMF. Although the results obtained by this approach are slightly degraded, the proposed algorithm can be applied to more general range.

In addition, R_i, mentioned in Table 1, is a 2×2 matrix, which is represented as follows:

$$S_i = R_i * \left[\widehat{s}_i ; \widehat{O}_i\right], i = 1, 2, \cdots, N, \tag{8}$$

where $\widehat{s}_i\widehat{O}_i$ denotes the speech signal of the target speaker and $\widehat{O}_i$ denotes the set of speech signals obtained after the separation of all speakers.

According to the defects of some existing single-channel speech separation methods, a new algorithm combining CNMF and JADE is proposed in this study. The CNMF is first trained using the training speech signal, and the trained CNMF is used to separate the mixed speech. Next, the separated speech signals are mixed, and the secondary separation is conducted by using JADE. In the next section, simulation experiments are performed to verify the performance of the proposed algorithm and compare it with several other algorithms.

3.2. Speech Enhancement. Some noise usually remains in the target speaker's speech after speech separation, and the interference of noise will inevitably reduce the quality and intelligibility of speech. For the mentioned reason, suppressing the background noise and extracting the pure speech becomes an important part of the speech processing process. Speech enhancement techniques should be used to enhance the target signal after speech signal separation. The conventional single-channel speech enhancement techniques comprise checkpoints [57], Wiener filtering [58], Kalman filtering [59], wavelet transform [60], and so on.

However, as reported by some existing studies, wavelet transform has more significant advantages in single-channel speech signal enhancement. Moreover, the experiments in this study prove this point. Wavelet transform is another landmark technique after Fourier transform. Wavelet transform inherits the advantages of Fourier transform while overcoming its defects. It is an ideal tool for signal time-frequency analysis and processing. One of the features of the wavelet transform in signal processing is that the transform can make certain aspects of the signal more prominent, so it is enabled to highlight signal details when processing the signal and thus extract the effective signal.

Accordingly, based on the above motivation, we will use wavelet transform as the speech signal enhancement technique in this study and propose a more effective adaptive wavelet transform to enhance the extracted signal.

In the following, the wavelet transform and the adaptive wavelet transform technique proposed in this study are introduced.

3.2.1. Speech Enhancement Based on Wavelet Transform. In the present section, we introduce the wavelet transform to enhance the sensor speech signal. The principle of wavelet transform is described below.

Set $L^2(R)$ as a square integrable space, and $\phi(t) \in L^2(t)$, if its Fourier transform satisfies Eq. (9) as follows:

$$C_\phi = \int_R \frac{|\phi(\omega)|^2}{|\omega|} d\omega < \infty. \tag{9}$$

$\phi(\omega)$ denotes a basic wavelet or a mother wavelet.

Table 1: Simulate the voice signal data acquired in different scenarios.

Scene	Number	Speaker	Target
2 speakers	a	1 female +1 female	
	b	1 female +1 male	
	c	1 male +1 male	
3 sparkers	d	1 female +2 males	1 female
	e	3 females	1 female
	f	2 females + male	1 female
	g	3 males	1 male

After the mother wavelet $\phi(t)$ is scaled and translated by a real pair (a, b), where $a, b \in R, a \neq 0$, a cluster function can be yielded:

$$\phi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\phi\left(\frac{t-b}{a}\right), a, b \in R\,; a \neq 0. \tag{10}$$

This cluster function denotes a wavelet basis function, where a represents the scaling factor and b denotes the translation factor. $\phi((t-b)/a)$ represents a window function whose window size is fixed but its shape can be changed. According to this characteristic, the wavelet transform is characterized by multiresolution analysis. $1/\sqrt{|a|}$ is a normalization factor, so the wavelets are enabled to have the same energy at different scales.

Signal processing based on wavelet domain is one of the main methods of speech signal processing. Wavelet transform has the characteristics of multiresolution, low entropy, and decorrelation, enabling the wavelet transform to show significant advantages in speech signals processing. Moreover, considerable wavelet bases can theoretically handle different scenarios, so the wavelet transform is significantly useful for speech signal processing.

The main process of wavelet transform denoising is shown in Figure 1, which well demonstrates the process.

3.2.2. Speech Enhancement Based on Adaptive Wavelet Transform. As suggested from the results of the experiments of this study, the quality of the enhanced speech signal may be reduced when the speech signal is enhanced after speech separation. This result proves that the speech enhancement algorithm cannot denoise properly on all noisy speech. In the present section, this study presents an adaptive method based on wavelet transform to analyze the CNMF+JADE separated speech signals and try to achieve selective speech enhancement, that is, before speech enhancement, automatic filtering those speech segments may cause quality degradation. As indicated from the analysis of the speech signal after separation and the speech after wavelet transform, under the significant difference between the separated speeches, the quality will reduce while increase with the wavelet transform. Based on the mentioned findings, the following method is developed to process adaptive judgment before speech enhancement.

$$\begin{aligned} \text{is\_enhance} &= \begin{cases} 1, & \text{disp}(\hat{s}_i, \hat{O}_i) \leq p * \left(\text{disp}(\hat{s}_i, \hat{O}_{i-1}) + \text{disp}(\hat{O}_i, \hat{O}_{i-1})\right)/2 \\ 0, & \text{otherwise}, \end{cases} \\ i &= 1, 2, \cdots, N, \\ O_{i-1} &= O_i + s_i + l, \\ O_0 &= O, \\ O_N &= s_N, \end{aligned} \tag{11}$$

where s_i denotes the ith target speaker speech signal after CNMF+JADE separation. O_i is the mixed speech signal after the CNMF+JADE separation on the mixed speech O_{i-1}. $\hat{s}_i$ and $\hat{O}_i$, respectively, express the Gaussian Mixture Model (GMM) [61, 62] of s_i and O_i. l indicates the loss during the separation process. N represents the number of speakers included in the mixed signal. p is the scaling factor, and the value is [1, 1.2].

disp$(\cdot)$ represents the GMM distance calculation formula, as defined below:

$$\text{disp}(A, B) = \sum_{i=1}^{M} WA_i \left(\sum_{j=1}^{M} WB_j d_{AB}(i, j)\right). \tag{12}$$

The function of disp$(\cdot)$ is to measure the dispersion between A and B, i.e., the coupling degree, and W is the weight.

Equation (11) can be explained as under the low coupling between and obtained by CNMF+JADE separation, no further speech enhancement is performed. In other words, under the 0 value obtained from Eq. (11), it is considered that the better the separation effect of the CNMF+JADE algorithm, the less noise the separated speech will contain, and then, further speech enhancement may be counterproductive. Furthermore, under the value of 1, the experimental wavelet transform is considered to be required for separation again.

Equation (11) adaptively determines which separated signals should be enhanced again and which ones do not, so the separated speech signals can be effectively optimized.

Finally, Figure 2 illustrates the flow of the whole algorithm.

4. Experiment Verification

As impacted by the limitations of the experimental conditions, in the present section, a sensor will be simulated to acquire speech data in a speech scene. The basic data used in the experiments originate from an acoustic-phonetic continuous speech corpus constructed in collaboration with Texas Instruments, MIT, and SRI International, i.e., the TIMIT dataset. The TIMIT dataset exhibits a speech sampling frequency of 16 kHz and comprises a total of 6300 sentences spoken by 630 individuals from eight major dialect regions in the United States. All sentences were manually segmented at the phoneme level (phone level) and then labeled. 70% of the speakers were male, and the speakers were

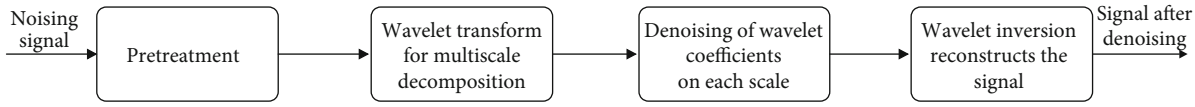


Figure 1: Wavelet denoising process diagram.

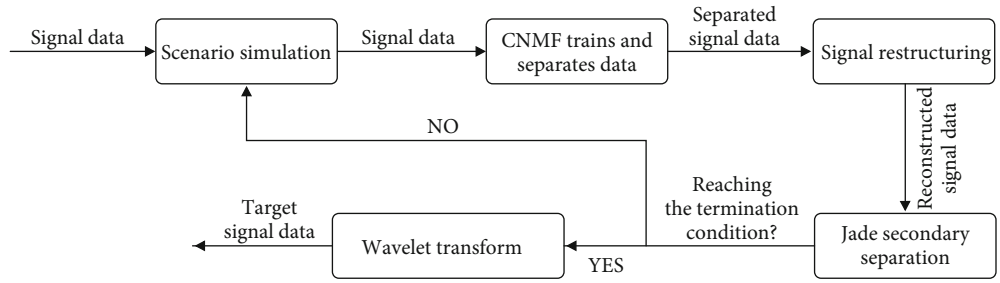


Figure 2: Flow chart of CNMF+JADE+wavelet transform.

primarily white adults. Next, multiple scenarios are simulated, and the speech signals are mixed according to the different scenarios.

For experimental design, in the first part of the experiment, different algorithms are used to separate the speech signals, and then, the signals are analyzed and compared with the algorithm proposed in this study to show that the CNMF +JADE algorithm proposed here can apply to the analysis and processing of the signal data collected by speech sensors. Subsequently, in the second part of the experiments, the performance of several single-channel speech enhancement techniques is verified, and the ability of the adaptive wavelet transform technique proposed in this study to effectively enhance the separated speech signals is experimentally verified, proving the effectiveness of the proposed method. In the following, the experiments are elucidated.

According to Table 1, the case of a speech was simulated, and two scenarios were set up. Scenario I contains two speakers and sets three specific scenarios. Scenario II contains three speakers, one of whom is the target speaker, and sets four specific scenarios. All the scenarios are set up with numbers *a*-*g*.

In addition, three scientific evaluation metrics are adopted to scientifically evaluate the quality of the separated speech signal. The three evaluation metrics introduced and their descriptions are elucidated below:

(1) Signal-Noise Ratio (SNR) [63] is the ratio between the valid signal and the invalid signal (noise signal). The larger the ratio, the greater the proportion of valid signals will be, and the purer the signal will be

(2) Perceptual Evaluation of Speech Quality (PESQ) [64] is an objective, full-reference speech quality assessment method that considers the subjective perception of human speech signals and can provide a subjective predictive value for objective speech quality assessment, which is recognized as an objective reflection of subjective evaluation. The PESQ score ranges between [-0.5,4.5], and a higher score indicates better speech quality after separation

(3) Short-Time Objective Intelligibility (STOI) [65], like PESQ, refers to a common objective evaluation method that conforms to the human auditory system for speech quality evaluation. It represents the actual intelligibility of speech, with the value ranging between [0,1]. If the value is closer to 1, the more easily the separated speech will be understood, and the higher the intelligibility will be

4.1. Speaker Separation. In the present section, simulation experiments are performed to verify the effectiveness of the proposed algorithm. According to the way of sound mixing, the speech signal separation falls to mono and multichannel speech separation. Since multichannel speech signals involve more available knowledge than monophonic speech signals, multichannel speech signals are simpler to process. The common multichannel speech separation algorithms are mainly based on Independent Components Analysis (ICA) and have shown better performance. For this reason, in the present section of experiments, we selected ICA as the comparison algorithm for speech separation. However, it should be noted that our simulation experiments are based on single-sensor hybrid speech separation, which does not satisfy the application of the ICA algorithm. Thus, in this part of the experiments, we extend the ICA algorithm by combining ICA with CNMF so that it can be applied to but-channel speech separation and compare it with the algorithm proposed in this study. Lastly, the specific methods used in this study are CNMF, CNMF+ICA, and CNMF+JADE.

Table 2 shows the results of the experiments by employing different separation methods, *a*-*g* corresponding to several dialogue scenarios simulated above in turn. The values in the table represent the evaluated results of the target speaker's speech and the original pure speech with the corresponding methods, in which the data corresponding to the MIX method refer to the data of the three metrics corresponding to the original mixed pure speech, and the later data are the results achieved with the three methods CNMF, CNMF+ICA, and CNMF+JADE, respectively. The best experimental results in each scenario are marked in italics.

Table 2: The results of different speech separation methods (SNR, PESQ, STOI).

Method	Index	Voice a	b	c	d	e	f	g
MIX	SNR	1.64	5.45	-0.77	1.83	-0.61	0.08	-2.07
	PESQ	2.27	2.05	2.43	1.58	1.84	1.75	1.73
	STOI	0.87	0.87	0.74	0.76	0.75	0.78	0.60
CNMF	SNR	9.31	7.93	8.46	*8.52*	5.94	6.38	4.89
	PESQ	2.85	2.21	2.49	1.85	1.99	1.97	2.08
	STOI	0.85	0.85	0.77	0.80	0.78	0.78	0.73
CNMF + ICA	SNR	9.28	10.62	5.42	6.88	*7.19*	5.71	2.63
	PESQ	2.10	1.84	1.80	1.78	1.80	1.66	1.78
	STOI	0.94	*0.93*	0.89	*0.89*	0.85	0.88	0.62
CNMF + JADE	SNR	*13.10*	*9.20*	*11.19*	8.44	7.03	*8.32*	*5.68*
	PESQ	*3.02*	*2.40*	*2.69*	*2.05*	*2.20*	*2.26*	*2.35*
	STOI	*0.95*	0.90	*0.92*	0.88	*0.85*	*0.89*	*0.74*

From the experimental results in the table, we can find that the speech signals processed by all methods are significantly improved compared to the original mixed speech MIX. In addition, the CNMF+JADE algorithm proposed in this study achieves the best experimental results in almost all scenarios; among the 7 scenarios and 21 metrics, only 4 metrics are worse than the experimental results of other methods (CNMF+ICA), which are the SNR and STOI results of scenario *b* and STOI results of scenario *d*. Moreover, it can be seen that the experimental results evaluated using PESQ are all better than those calculated by several other algorithms, which fully demonstrates the effectiveness of the proposed algorithm.

First, the proposed CNMF+JADE algorithm is compared with the CNMF algorithm, and all experimental results are found to outperform those of CNMF, which demonstrates that combining JADE with CNMF is effective. Subsequently, as revealed from the comparison with the CNMF+ICA algorithm, almost all the results are better than those achieved by CNMF+ICA, indicating that combining JADE with CNMF is a purposeful combination and more promising. The combined experimental results fully illustrate the effectiveness of the proposed algorithm.

4.2. The First Experiment Verification for Enhancement. In this part of the experiments, the performance of several conventional single-channel speech signal enhancement techniques is compared. The separated signal complies with the signal of the target speaker obtained from the CNMF +JADE method in Section 4.1. Moreover, the CNMF+JADE method is the method proposed in this study. Subsequently, the target speech signal is enhanced with the four speech enhancement methods separately, and lastly, the enhanced speech signal is evaluated with SNR, PRSQ, and STOI. The experimentally achieved results are listed in Table 3, where the experimental results of the CNMF+JADE method represent the experimental results to be compared. Likewise, *a-g* columns correspond to the various scenarios in Table 1, in which each method is evaluated with three evaluation metrics.

First, comparing the four conventional single-channel speech enhancement methods, it can be found that the algorithm using wavelet transform as the speech enhancement method exhibits the optimal performance among the four conventional speech enhancement methods. As suggested by the experimental results achieved with SNR as the evaluation index, the wavelet transform achieves the optimal results in all seven scenarios. For the experimental results achieved with STOI as the evaluation index, six scenes also achieve the optimal results, and only the experimental results of scenario *a* are slightly lower than those of the wiener filtering method, and the differences are slight, 0.82 and 0.83, respectively. Specific to the experimental results achieved with PESQ as the evaluation index, four of the seven scenes achieve the optimal results. As indicated from the comprehensive experimental results, the enhancement of the speech signal obtained by separating CNMF+JADE algorithm using wavelet transform is very effective. For the mentioned reason, this is one of the motivations for choosing wavelet transform as the speech enhancement method in this study.

In addition, the results of the experiments in which the wavelet transform method is used are compared with the results of the experiments in which the speech enhancement method is not used. It can be found that not all the speech quality is enhanced after speech enhancement. For instance, specific to scenario *a*, the speech quality obtained after using the wavelet transform method decreases in all cases. For the experimental results achieved by using wavelet transform as the speech enhancement method, a total of 11 results out of 7 scenes and 21 results are better than the experimentally achieved results without the speech enhancement method.

For this reason, it can be concluded that the purpose of speech enhancement is to remove the noise in the speech segment and thus improve the quality of speech. However, during speech enhancement, the speech signal is corrupted to a certain extent, so the speech quality turns out to be not necessarily better after speech enhancement.

Table 3: The results of the enhancement methods (SNR, PRSQ, STOI).

Method	Index	Scene *a*	*b*	*c*	*d*	*e*	*f*	*g*
CNMF + JADE	*SNR*	13.10	9.20	11.19	8.44	7.03	8.32	5.68
	PESQ	3.02	2.40	2.69	2.05	2.20	2.26	2.35
	STOI	0.95	0.90	0.92	0.88	0.85	0.89	0.74
CNMF + JADE+ spectral subtraction	*SNR*	5.22	5.36	6.16	4.76	4.07	4.76	3.55
	PESQ	1.97	1.95	1.70	1.89	1.88	1.89	1.27
	STOI	0.79	0.82	0.70	0.79	0.72	0.80	0.75
CNMF + JADE+ Wiener filtering	*SNR*	-2.24	-2.04	-1.02	-1.96	-2.13	-1.96	-1.62
	PESQ	2.36	2.33	2.96	2.13	2.19	2.13	2.49
	STOI	0.73	0.77	0.72	0.72	0.69	0.72	0.72
CNMF + JADE+ Kalman filtering	*SNR*	3.27	3.42	3.19	3.05	2.84	4.0	2.05
	PESQ	2.03	2.25	2.05	2.00	1.97	2.00	2.04
	STOI	0.83	0.88	0.69	0.82	0.78	0.83	0.57
CNMF + JADE+ wavelet transform	*SNR*	12.02	11.10	15.9	8.35	7.23	8.37	7.40
	PESQ	2.49	2.46	1.96	1.70	2.25	2.15	1.90
	STOI	0.82	0.92	0.81	0.84	0.88	0.91	0.76

Thus, it is very important and necessary to adaptively select the speech signals that should be enhanced, instead of blindly enhancing all signals. For this reason, this study proposes an adaptive wavelet transform method that adaptively selects the enhanced speech signals and filters out the speech signals that are not required to be enhanced. The specific experimental validation is presented in the next section.

4.3. The Second Experiment Verification for Enhancement. In this part of the experiments, the adaptive wavelet transform enhancement method proposed in this study is validated. Again, the enhanced speech signal is acquired from the speech signal obtained after separation using the CNMF +JADE method. Moreover, the experimental results of the three metrics are verified separately. The achieved experimental results are listed in Tables 4–6, which fall to three parts, i.e., CNMF+JADE for the experimental results without enhancement and CNMF+JADE+wavelet transform for the experimental results with wavelet transform. Lastly, the adaptive wavelet transform method proposed here is adopted to evaluate whether the speech signal should be enhanced in each scene. From the experimental results in Tables 4–6, we can see that 0 is the experimental result without enhancement, and the corresponding experimental results with wavelet transform enhancement have decreased. 1 is the experimental result with enhancement, and the corresponding experimental results with wavelet transform enhancement have improved.

It is demonstrated through experiments that our adaptive judgment method can filter out the speech segments whose quality will be degraded after wavelet transform. As revealed from the results, the adaptive wavelet transform speech enhancement method proposed in this study can automatically filter the speech segments that are not suitable for speech enhancement, thus effectively improving the quality of the final speech signal.

4.4. Compared with the Deep Learning. In recent years, with the development of deep learning, researchers have noticed that the nonlinear processing and feature learning capabilities of deep models have significant advantages in addressing speech separation problems. Thus, in this part of the experiments, we implemented a cyclic stacking neural network (Ref [66]) to perform separation processing of the acquired speech signals. In Ref, the speech separation results of various deep neural networks are compared, which are close to the work in this study. We use two metrics, PESQ and STOI, to evaluate the quality of the separated speech signal to compare the performance of the proposed algorithm with deep learning algorithms. Comparing the results of the proposed speech separation methods, we can dig out the advantages and disadvantages of the shallow and deep models.

The experimental results of the proposed algorithm and the deep learning algorithm are shown in Table 7. From the experimental results, we can see that there is still a gap between the method proposed in this study and the deep learning method. In terms of PESQ index, the improvement of RDSN is obviously better than the method in this study. As indicated from the experimental results achieved with STOI as the evaluation index, the optimal value of the proposed method in this study is 0.106, which is the same as the experimental result of DDN, and the difference with the experimental result of RDSN is not much, only 0.006.

As indicated from a comprehensive analysis of the experimental results, the deep model outperforms the shallow model in the supervised case. However, the deep model requires considerable training data, and a large amount of speech data are very difficult to obtain. In addition, the deep model is more expensive to train, and it is difficult to achieve

Table 4: The experiment of adaptive judgment speech enhancement (SNR).

Method	Scene						
	a	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	13.10	9.20	11.19	8.44	7.03	8.32	5.68
CNMF + JADE+ wavelet transform	12.02	11.10	15.9	8.35	7.23	8.37	7.4
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 5: The experiment of adaptive judgment speech enhancement (PESQ).

Method	Scene						
	a	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	3.02	2.40	2.69	2.05	2.20	2.26	2.35
CNMF + JADE+ wavelet transform	2.49	2.46	1.96	1.70	2.25	2.15	1.90
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 6: The experiment of adaptive judgment speech enhancement (STOI).

Method	Scene						
	a	***b***	***c***	***d***	***e***	***f***	***g***
CNMF + JADE	0.95	0.90	0.92	0.88	0.85	0.89	0.74
CNMF + JADE+ wavelet transform	0.82	0.92	0.81	0.84	0.88	0.91	0.76
Adaptive speech enhancement judgment	0	1	1	0	1	1	1

Table 7: Efficiency comparison of speech separation effect.

Method	Index	
	PESQ	*STOI*
Deep neural networks (*DNN*) [55]	0.694	0.106
Recurrent deep stacking networks (*RDSN*) [55]	0.823	0.112
CNMF + JADE+ adaptive wavelet transform	0.305	0.106

small-sample, unsupervised speech separation in complex scenarios. The speech separation algorithm proposed in this study can satisfy the needs of small sample and unsupervised speech separation. In addition, the total computational overhead of the shallow model is smaller than that of the deep model. As opposed to the deep model, the shallow model is more suitable for application scenarios with high real-time requirements. Given the comparison of the two models synthetically, the algorithm proposed in this study is considered to be more suitable for target speaker speech extraction in the complex multispeaker scenario.

5. Conclusion

The development of IoT technology promotes the rapid development of intelligent voice systems, and the efficient processing of signal data acquired by speech sensors becomes imminent. Thus, an unsupervised speech separation algorithm based on the combination of CNMF and JADE is proposed in this study. Through simulation experiments, it is well demonstrated that the proposed algorithm can effectively separate the target speech signals contained in the mixed speech signals. In addition, for the separated speech signal is weak and out of frame, this study also proposes an adaptive wavelet transform method to enhance the separated speech signal. As revealed from the results, the proposed algorithm in this study can enhance the separated speech signals. The comprehensive experimental results can prove that the proposed algorithm is very competitive in the processing of single-channel mixed speech separation problem. The algorithm is highly versatile and robust, capable of technically supporting other researchers in processing highly noisy signal data collected by sensors.

Speech separation, especially single-channel speech separation, has been a hotspot and difficult research area. In addition, as IoT technology is being developed and applied, separating high-quality speech signals has become an urgent task. Speech signals exhibit obvious spatio-temporal structures and nonlinear relationships, and most of the conventional speech classification methods are shallow structures, and the mentioned results are more limited in their ability to tap into the mentioned nonlinear structural information. In recent years, as deep learning is advancing, it has been suggested that the nonlinear processing and feature learning capabilities of deep models exhibit obvious advantages in addressing speech separation problems. Moreover, some results of processing speech signals with deep learning have been published. As deep learning computing is leaping forward, deep models (e.g., DNN, DSN, CNN, RNN, Deep NMF, and LSTM) will definitely be more competitive in speech separation problems. In the future, the use of deep learning techniques in speech separation will definitely become a research hotspot.

Data Availability

We are using the TIMIT dataset, which can be found at https://academictorrents.com/details/34e2b78745138186976cbc27939b1b34d18bd5b3/techamp; hit =1amp; filelist =1.

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 (61902232, 61902231), the Natural Science Foundation of Guangdong Province (2019A1515010943), the Key Project of Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Natural Science) (2018KZDXM035), the Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence) (2019KZDZX1030), and the 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D).

References

[1] S. S. Hao, Y. P. Wang, and B. C. Lv, "A new optimisation model and algorithm for virtual optical networks," *International Journal of Sensor Networks*, vol. 29, no. 4, pp. 252–261, 2019.

[2] C. X. Ji, Y. P. Wang, Z. Q. Xu, and X. Li, "A new model and algorithm for RSA problem in elastic optical networks," *International Journal of Sensor Networks*, vol. 31, no. 3, pp. 145–155, 2019.

[3] M. Ye, Y. Wang, C. Dai, and X. Wang, "A hybrid genetic algorithm for the minimum exposure path problem of wireless sensor networks based on a numerical functional extreme model," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8644–8657, 2016.

[4] H. Y. Liu, Y. P. Wang, and N. L. Fan, "A hybrid deep grouping algorithm for large scale global optimization," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 6, pp. 1112–1124, 2020.

[5] Y. Wang, H. Liu, F. Wei, T. Zong, and X. Li, "Cooperative coevolution with formula-based variable grouping for large-scale global optimization," *Evolutionary Computation*, vol. 26, no. 4, pp. 569–596, 2018.

[6] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1266–1279, 2017.

[7] S. Srinivasan and D. L. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Communication*, vol. 52, no. 1, pp. 72–81, 2010.

[8] L. Ferrer, M. K. Nandwana, M. Mclaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: trial-based calibration with a reject option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2018.

[9] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, "A blind source separation techinique based on second order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 2002.

[10] H. Zhang, G. Hua, L. Yu, Y. Cai, and G. Bi, "Underdetermined blind separation of overlapped speech mixtures in time-frequency domain with estimated number of sources," *Speech Communication*, vol. 89, pp. 1–16, 2017.

[11] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 989–1001, 2011.

[12] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 3, pp. 662–675, 2013.

[13] N. Tengtrairat, W. L. Woo, S. S. Dlay, and B. Gao, "Online noisy single-channel source separation by adaptive spectrum amplitude estimator and masking," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1881–1895, 2016.

[14] Z. Fan, Y. Lai, and J. R. Jang, "SVSGAN: singing voice separation via generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 726–730, Calgary, AB, Canada, April 2018.

[15] M. Michelashvili, S. Benaim, and L. Wolf, "Semi-supervised monaural singing voice separation with a masking network trained on synthetic mixtures," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295, Brighton, UK, May 2019.

[16] X. Sun, J. Xu, Y. Ma, T. Zhao, and S. Ou, "Single-channel blind source separation based on attentional generative adversarial network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, pp. 1–8, 2020.

[17] L. Xie, L. Tan, and M. W. Mak, "Guest editorial: advances in deep learning for speech processing," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 1–3, 2018.

[18] T. Ogunfunmi, R. P. Ramachandran, R. Togneri, Y. Zhao, and X. Xia, "A primer on deep learning architectures and applications in speech processing," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3406–3432, 2019.

[19] T. Zhao, Y. Zhao, and C. Xin, "Ensemble acoustic modeling for CD-DNN-HMM using random forests of phonetic decision trees," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 187–196, 2016.

[20] Z. Yan, H. Qiang, and X. Jian, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR," *Mathematics of Computation*, vol. 44, no. 170, pp. 519–521, 2013.

[21] Y. Wang, J. Du, L. R. Dai, and C. H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[22] J. Du, Y. Tu, L.-R. Dai, and C. H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2017.

[23] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proceedings of the International Joint Conference on Neural Networks, 2003*, Portland, OR, USA, July 2003.

[24] B. Chen, G. Polatkan, G. Sapiro, D. Blei, D. Dunson, and L. Carin, "Deep learning with hierarchical convolutional factor analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1887–1901, 2013.

[25] D. T. Pham and J. F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[26] C. Vaz, D. Dimitriadis, S. Thomas, and S. Narayanan, "CNMF-based acoustic features for noise-robust ASR," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,, pp. 5735–5739, Shanghai, China, March 2016.

[27] Y. Zhang, S. Qi, and L. Zhou, "Single channel blind source separation for wind turbine aeroacoustics signals based on variational mode decomposition," *IEEE Access*, vol. 6, pp. 73952–73964, 2018.

[28] Z. Yan, S. Sheng-kai, L. Yue, and W. Jia-qi, "Sonar echo signal processing based on Convolution Blind source separation," in *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 130–134, Shanghai, China, September 2020.

[29] B. Wiem, B. M. Mohamed Anouar, P. Mowlaee, and B. Aicha, "Unsupervised single channel speech separation based on optimized subspace separation," *Speech Communication*, vol. 96, pp. 93–101, 2017.

[30] S. Mavaddaty, S. M. Ahadi, and S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Communication*, vol. 76, pp. 42–60, 2015.

[31] Q. H. Lin, F. L. Yin, T. M. Mei, and H. Liang, "A blind source separation based method for speech encryption," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 6, pp. 1320–1328, 2006.

[32] H. T. Fan, J. W. Hung, X. Lu, S.-S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[33] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, & Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[34] T. Pham, Y. S. Lee, Y. A. Chen, and J.-C. Wang, "A review on speech separation using NMF and its extensions," in *2015 International Conference on Orange Technologies (ICOT)*, pp. 26–29, Hong Kong, China, December 2016.

[35] S. Lee, D. Han, and H. Ko, "Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities," *Applied Acoustics*, vol. 117, pp. 257–262, 2017.

[36] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, Corfu, Greece, July 2011.

[37] T. Pham, Y. S. Lee, Y. B. Lin, T.-C. Tai, and J. Wang, "Single channel source separation using sparse NMF and graph regularization," in *ASE BD&SI '15: Proceedings of the ASE BigData & SocialInformatics 2015*, p. 55, New York, NY, USA, October 2015.

[38] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[39] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: how models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[40] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, vol. 24, no. 1, pp. 16–29, 2010.

[41] P. Mowlaee, R. Saeidi, M. G. Christensen et al., "A joint approach for single-channel speaker identification and speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2586–2601, 2012.

[42] R. Saeidi, P. Mowlaee, T. Kinnunen et al., "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, August 2010.

[43] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2011.

[44] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[45] H. Zhang, X. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm for pitch estimation and speech separation using deep stacking network," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 246–250, South Brisbane, QLD, Australia, April 2015.

[46] J. Chang and D. L. Wang, "Robust speaker recognition based on DNN/i-vectors and speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5415–5419, New Orleans, LA, USA, March 2017.

[47] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, South Brisbane, QLD, Australia, April 2015.

[48] S. Wisdom, T. Powers, J. Pitton, and L. Atlas, "Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.

[49] F. Weninger, H. Erdogan, S. Watanabe et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation. LVA/ICA 2015. Lecture Notes in Computer Science, vol 9237*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds., pp. 91–99, Springer, Cham, 2015.

[50] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[51] Y. H. Tu, J. Du, and C. H. Lee, "A speaker-dependent approach to single-channel joint speech separation and acoustic modeling based on deep neural networks for robust recognition of multi-talker speech," *Journal of Signal Processing Systems*, vol. 90, pp. 963–973, 2017.

[52] D. Yu, M. Kolbk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent

multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, New Orleans, LA, USA, March 2017.

[53] M. Kolbk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[54] Y. Wang, J. Du, L. R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *Interspeech 2017*, pp. 1178–1182, Stockholm, Sweden, August 2017.

[55] Z. Q. Wang and D. L. Wang, "Recurrent deep stacking networks for supervised speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75, New Orleans, LA, USA, March 2017.

[56] N. Dey and A. S. Ashour, "Applied examples and applications of localization and tracking problem of multiple speech sources," in *Direction of Arrival Estimation and Localization of Multi-Speech Sources. SpringerBriefs in Electrical and Computer Engineering*, pp. 35–48, Springer, Cham, 2018.

[57] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, 2002.

[58] M. Joham, K. Kusume, M. H. Gzara, W. Utschick, and J. A. Nossek, "Transmit Wiener filter for the downlink of TDDDS-CDMA systems," in *IEEE Seventh International Symposium on Spread Spectrum Techniques and Applications*, pp. 9–13, Prague, Czech Republic, September 2002.

[59] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 191–195, Shanghai, China, March 2016.

[60] T. Gülzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, pp. 5–19, 1998.

[61] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[62] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[63] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *Journal of Speech and Hearing Research*, vol. 29, no. 2, pp. 146–154, 1986.

[64] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Salt Lake City, UT, USA, May 2002.

[65] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, March 2010.

[66] R. Takashima, Y. Kawaguchi, Q. Sun, T. Sumiyoshi, and M. Togami, "An application of noise-robust speech translation using asynchronous smart devices," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1592–1595, Kuala Lumpur, Malaysia, December 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6651112, 9 pages
https://doi.org/10.1155/2021/6651112

Research Article

Energy Efficiency Opposition-Based Learning and Brain Storm Optimization for VNF-SC Deployment in IoT

Hejun Xuan,[1] **Xuelin Zhao**,[1] **Zhenghui Liu**,[1,2] **Jianwei Fan**,[1,2] **and Yanling Li**[1,2]

[1]*School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China*
[2]*Henan Key Lab. of Analysis and Application of Education Big Data, Xinyang Normal University, Xinyang 464000, China*

Correspondence should be addressed to Hejun Xuan; xuanhejun0896@xynu.edu.cn

Received 25 November 2020; Revised 30 December 2020; Accepted 3 February 2021; Published 13 February 2021

Academic Editor: Pei-Wei Tsai

Network Function Virtualization (NFV) can provide the resource according to the request and can improve the flexibility of the network. It has become the key technology of the Internet of Things (IoT). Resource scheduling for the virtual network function service chain (VNF-SC) is the key issue of the NFV. Energy consumption is an important indicator for the IoT; we take the energy consumption into the objective and define a novel objective to satisfying different objectives of the decision-maker. Due to the complexity of VNF-SC deployment problem, through taking into consideration of the heterogeneity of nodes (each node only can provide some specific VNFs), and the limitation of resources in each node, a novel optimal model is constructed to define the problem of VNF-SC deployment problem. To solve the optimization model effectively, a weighted center opposition-based learning is introduced to brainstorm optimization to find the optimal solution (OBLBSO). To show the efficiency of the proposed algorithm, numerous of simulation experiments have been conducted. Experimental results indicate that OBLBSO can improve the accuracy of the solution than compared algorithm.

1. Introduction

Internet of Things (IoT) is turning into the future generation of wireless network communication technology and sensor networks. IoT devices are cost-effective, so distributed expensive devoted spectrum to the IoT would be inefficient. Therefore, the problem of resource allocation that satisfies the constraints of network operation is particularly important. The IoT can work as an additional layer on the basic network of any communication technology [1, 2]. The spectrum access problem solution for the IoT network is a combination of low-cost networks that can be combined. The use of a variety of network technologies is to serve the traffic of the IoT [3]. Eternal virtualization has many advantages, which can improve VONs with different widely used topologies. In addition, it can enable VONs to share physical network explorer between different users and applications, reduce physical resource management, and provide simple spectrum allocation [4–7]. However, how to map a large number of von of different topologies to the physical network while achieving certain goals, such as energy consumption, blocking rate, and network performance, is a challenge [8, 9]. In recent years, there has been a lot of research focused on the VONs mapping problem and related issues [10]. Virtualization (NFV) with Network Function Virtualization makes it possible to manage mobility within the infrastructure such as the Service Function Chaining (SFC). With the change of Network load and node and link state in the infrastructure, the migration of Virtual Network Function (VNF) in SFC can improve the utilization rate of underlying resources and meet the requirements of different slices for delay. In addition, the flexible arrangement of VNF also provides a favorable condition for saving system energy consumption. Under the mechanism of VNF sharing, VNF migrates servers with low resource utilization, and shutdown of the corresponding server can achieve the purpose of reducing energy consumption [11, 12].

In this work, we studied the spectrum allocation of VON mapping service chain (VNF-SC) in the IoT. Different from the previous work, we have studied the IoT, that is, each node

can only provide some specific VNFs, and the system resources of all nodes are limited. The major contributions of this study are summarized as follows:

(i) Energy consumption is an important indicator for the IoT; we take the energy consumption into the objective and define a novel objective to satisfying different objectives of the decision-maker

(ii) Due to the complexity of VNF-SC deployment problem, through taking into consideration of the heterogeneity of nodes (each node only can provide some specific VNFs), and the limitation of resources in each node, a novel optimal model is constructed to define the problem of VNF-SC deployment problem

(iii) Since a large number of variables are in the optimal model, it has numerous local optimal solutions. For the sake of jumping out the local optimal, a weighted center opposition learning strategy is proposed. Based on this, an improved brain storm optimization algorithm, which can improve the accuracy of the solution, is proposed. Experimental results demonstrate that the proposed algorithm can obtain a better solution than the compared algorithm

2. Related Work

Last few years, some studies have been conducted on network scheduling issues using VNF-scs, mainly focusing on service link routing and VNF deployment issues (e.g., [13–15]). Literature [16] minimizes the sum of the three costs of cloud resource cost, bandwidth cost, and reconstruction cost. The characteristic of reconstruction cost is the loss of revenue generated by network operators due to bit loss. Considering the constraints of flexible optical network and DC capacity, an effective algorithm based on noncooperative mixed strategy game is proposed [17]. In order to solve the relatively long setup delay and complicated network control problems, we designed a configuration framework with resource predeployment to solve the above problems [18]. The proposed game model enables tenants to compete for VNF-SC supply services based on the incentives of income and service quality, so it can encourage tenants to choose more reasonable supply solutions. In order to meet the needs of users and maximize the benefits of suppliers, a VNF deployment algorithm based on eigenvalue decomposition is proposed [19]. So far, most of the literatures, such as literature [20], have studied the issues related to the migration of virtual machines under a certain migration trigger time, but such a migration strategy does not consider multiple SFC business scenarios at the same time. Literature [21] studies the VNF migration problem under the one-to-one mapping relationship between service functional chain VNF and node VNF instances. In the shared state of VNF instances, if the performance of a slice fails to meet user requirements, the current research cannot formulate an effective strategy to achieve VNF migration. Literature [22] proposed a VNF migration algorithm based on MDP theory to deal with the constantly changing workload, and its migration strategy was aimed at minimizing energy consumption and reconfiguration cost caused by VNF migration. Literature [23] established a cost model and proposed a greedy algorithm to optimize the migration of VNF, but this scheme can only solve the problem of resource allocation in a single scheduling cycle. Literature [24] makes backup for the whole SFC, which increases the resource overhead. In literature [25], a reliability perception method combining VNF deployment and routing optimization is proposed, which adopts a backup sharing method to reduce resource consumption. While using the backup mechanism to improve reliability, the link length of SFC is increased, and the end-to-end delay of SFC is increased to a certain extent. In literature [26], no backup mechanism was adopted. PageRank thought was adopted when deploying VNF, and reliability and delay were considered at the same time. However, when deploying VNF, the source node and destination node were not considered, so the delay was increased. Literature [27] did not adopt the backup mechanism and proposed the SFC mapping algorithm based on queue awareness to improve the stability and reliability of the network, which took into account the source nodes and destination nodes. Literature [27] does not constrain VNF types but assumes that each physical node can carry any type of VNF. Literature [27] pointed out that adjacent VNF in the same SFC can be aggregated, that is, deployed on the same physical node, but it did not provide a specific polymerization method.

3. Problem Description and Modeling

3.1. Network and VNF-SC Description. $G(V, E)$ is an undirected graph, and we use it to denote an IoT, where $V = \{v_1, v_2, \cdots, v_{N_V}\}$ is nodes set in the network. N_V denotes the number of nodes, v_i is the ith optical node. $E = \{l_{ij} \mid v_i, v_j \in V\}$ denotes the set of links. l_{ij} represents the link between v_i and v_j. N_E represents the number of links. Each link has N_F frequency slots (FSs), and the indexes of FSs on each link are $1, 2, \cdots, N_F$.

Each VNF-SC is a task. Now, we have a set of VNF-SCs, denoted by $T = \{T_1, T_2, \cdots, T_{N_T}\}$, where N_T represents the tasks (VNF-SCs) number, and T_k is the kth VNF-SC. T_k can be described as $T_k = (s_k, d_k, VNF_k^T, b_k)$, where s_k and d_k represent source node and destination node, and $VNF_k^T = \{VNF_{k_1}, VNF_{k_2}, \cdots, VNF_{k_{M_k}}\}$ is the of virtual network functions to be realized in T_k, and M_k denotes the number of virtual network functions in VNF_k^T, i.e., some nodes should be chosen in the selected path to realize these virtual network functions. $b_k = (b_k^0, b_k^1, \cdots, b_k^{M_k})$ represents the frequency slots numbers of T_k required, where b_k^0 is the original number of frequency slots occupied by T_k. What is more, $\forall VNF_k^T \subseteq VNF (1 \le k \le N_T)$.

4. Energy Model

The total energy consumption of telecommunications networks supporting virtual network functions consists of two

parts. Let N_f^v denotes the number of virtual network functions deployed on node v:

$$N_f^v = \left\lceil \frac{\sum_{r\in R} z_{r,f}^v \cdot b_r}{ct_f} \right\rceil. \quad (1)$$

Since server energy consumption is positively correlated with CPU utilization, the total energy consumption of VNF f deployed on node v can be derived as

$$p_f^v = \frac{p_h^s - p_b^s}{C_v} \cdot cr_f \cdot \frac{\sum_{r\in R} z_{r,f}^v \cdot b_r}{ct_f}, \forall v \in V, f \in F. \quad (2)$$

Among them, p_b^s represents the startup energy consumption of the node s, and p_h^s represents the peak load energy consumption of the node. Let p_v denote the energy consumption of node s:

$$p_v = p_b^s \cdot \min\left\{1, \sum_{f\in F}\sum_{r\in R} z_{r,f}^v\right\} + \sum_{f\in F} p_f^v, \quad (3)$$

where $\min\{1, \sum_{f\in F}\sum_{r\in R} z_{r,f}^v\} \in \{0,1\}$ means some VNF instance is deployed on the node v. When it is equal to 1, the node v must be opened to carry the VNF instance. Represents the link bandwidth utilization, let U_l denote the bandwidth utilization of link l:

$$U_l = \frac{\sum_{r\in R} w_r^{ufvg} \cdot y_r^{uvl} \cdot b_r}{C_l}, \forall l \in L, f \in F. \quad (4)$$

The energy consumption of the link can be calculated as follows:

$$p_l = p_b^l \cdot \min\left\{1, \sum_{f\in F}\sum_{r\in R} w_r^{ufvg} \cdot y_r^{uvl}\right\} + \left(p_h^l - p_b^l\right) \cdot U_l. \quad (5)$$

Among them, p_b^l is the start-up energy consumption of the link l, and p_b^l is the peak load energy consumption of the link. $\min\{1, \sum_{f\in F}\sum_{r\in R} w_r^{ufvg} \cdot y_r^{uvl}\} \in \{0,1\}$ represents the on/off status of the dollar link l; when it is equal to 1, the link l must be powered on to the VNF-SC. Therefore, the total energy consumption of NFV-SC can be calculated as

$$p_{\text{total}} = \sum_{v\in V} p_v + \sum_{l\in L} p_l. \quad (6)$$

The other two objectives are given in the [28]. The objective function is expressed by

$$\min H = \min\{\alpha_1 p_{\text{total}} + \alpha_2 f_2 + \alpha_3 f_3\}, \quad (7)$$

where f_1 and f_2 denotes the maximum index of used frequency slots and ratio of resource used. Some constraint conditions should be satisfied. These constraint conditions are given in our previous paper [29]. To solve the optimization model, we propose an improved brain storm optimization algorithm, and the algorithm will be described in the following sections.

5. Proposed Algorithm

5.1. Bounding-Box Determines the Search Area. The basic idea of the Bounding box algorithm is to determine the possible Bounding rectangular regions of unknown nodes by measuring their distance from unknown nodes and using their distance. Finally, the center of masses of all Bounding regions is taken as the estimated position of unknown nodes. Although the algorithm is simple and easy to implement, its positioning accuracy is relatively low. References [30] put forward individual Bounding box initiative-inspired optimization algorithm to improve convergence speed and avoid overturning ambiguity in the positioning process. However, due to the impact of ranging errors, the real position of unknown nodes may fall outside the Bounding box region, reducing convergence speed and solving accuracy. To avoid this phenomenon, this paper improves the Bounding box method and determines the search area through multiple signal measurements and compensated measurement distances. The individual heuristic algorithm is initialized in the search area to improve convergence speed and solution accuracy. Assuming that there are m anchor nodes in the communication range of the unknown node, the coordinates of the unknown node (x_i, y_i) satisfy equation (8):

$$\begin{cases} x_i \in \left[x_j - d_{ij}, x_j + d_{ij}\right] \\ y_i \in \left[y_j - d_{ij}, y_j + d_{ij}\right] \end{cases}. \quad (8)$$

5.2. Weighted Center Opposition Learning Strategy. The Opposition based Learning (OBL) strategy was proposed by Tizhoosh in 2005 [31]. It has been widely used in various algorithms, effectively improving the efficiency of solving the global optimum. The basic idea of the reverse learning strategy is as follows: in the search process, the initial position and its reverse position relative to the center are considered simultaneously, so as to enhance the diversity of individual groups and improve the global search capability of the algorithm. Let point $P = (p^1, p^2, \cdots, p^i, \cdots, p^d)$ is a point in D dimensional space, and $p_{\min}^i \le p^i \le p_{\max}^i (i = 1, 2, \cdots, d)$. $p_{\min}^i$ and $p_{\max}^i$ are the minimum and maximum value of point p^i in dimension D, respectively. Then, the opposite point of point p^i is

$$\widehat{p^i} = p_{\max}^i + p_{\min}^i - p^i. \quad (9)$$

Literature [32] proposed that the search population takes the mean mixing center as the symmetric center of reverse learning to guide the population evolution. Among them, the mean blending center is determined by the fitness value of the mean value of all individuals and the mean value of some better individuals. As the center of the mixed mean is located in the center of the group, individual positions can be integrated in the search process, which will promote the

1 Propose problems that need to be solved and gather people with different knowledge backgrounds;
2 These people come up with many solutions based on this problem and following the four principles;
3 Find several parties as owners of the problem and choose from each of these ideas the best solution they think will solve the problem;
4 Select some solutions to come up with more solutions, and the better solution in Step 3 has a greater probability of being selected;
5 Similar to step 3, the problem owner selects several better solutions;
6 Randomly select a solution and use its functions and characteristics as clues to generate more solutions;
7 Let the owner of the problem choose a few better solutions from all the solutions;
8 End up with a solution that's good enough;

Algorithm 1: Brainstorming Process.

group to draw closer to the center in the early stage and accelerate the convergence speed. After convergence, individuals can jump out of the optimal group and, thus, have a higher probability to search for the global optimal solution. Inspired by his ideas, we propose a weighted center reverse learning strategy. The basic idea is that the weight of all individuals in the population is given according to their fitness value, the weighted center of the population is calculated according to the weight and individual position, and the weighted center is taken as the symmetric center of reverse learning. Since fitness values of all individuals are taken into account, the weighted center can better reflect the central trend of the population in the current environment. Therefore, this paper uses the weighted center as the symmetric center of reverse learning to guide population evolution.

The fitness value $J(SP_i)$ of an individual is the value of the objective function. When solving the weight center, the individual weight is calculated as follows:

$$W_i = \frac{1/J(SP_i)}{\sum_{j=1}^{m}\left(1/J\left(SP_j\right)\right)}. \tag{10}$$

SP_i^d and $\bar{C}^d$, respectively, represent the values of individual I and the weighted center on dimension D. $\bar{C}^d$ can be calculated by

$$\bar{C}^d = \sum_{i=1}^{m}\left(W_i \cdot SP_i^d\right). \tag{11}$$

Opposition point position of point p on dimension D with the weighted center as the symmetric point can be calculated by

$$\bar{p}^d = k\left(\bar{C}^d - p_d\right) + \bar{C}^d, \tag{12}$$

where p_d is the value of point p on dimension d, $\bar{p}^d$ is the value of point p on dimension D after opposite learning by weighted center, and $k \in [0, 1]$ is the dynamic learning factor.

6. Brain Storm Optimization

Swarm intelligence optimization algorithm is an optimization algorithm inspired by the biological behavior of nature. It is considered as a young and promising swarm intelligence optimization algorithm, which simulates the brainstorm process of human beings in solving problems.

6.1. Brainstorming Process. When we come across a difficult problem that cannot be solved by one person alone, we will gather people with different knowledge backgrounds to brainstorm, and usually, the problem will be solved with a high probability. The specific steps of the brainstorming process are as follows:

In this paper, a weighted center opposition based learning is introduced to brainstorm optimization to find the optimal solution (OBLBSO), the improved Brain Storm Optimization is shown in Algorithm 2.

The brain storm optimization algorithm is a new one which simulates the brain storm conference process design. When solving the static single objective optimization problem, the detailed steps of the algorithm are shown in Algorithm 1. Static single objective brainstorm optimization algorithm mainly consists of three parts: individual clustering in decision space, generating new individuals, and selecting better individuals.

7. Experiments and Analysis

In order to verify the effectiveness and effectiveness of the algorithm, experiments are carried out on the NSFNET topology.

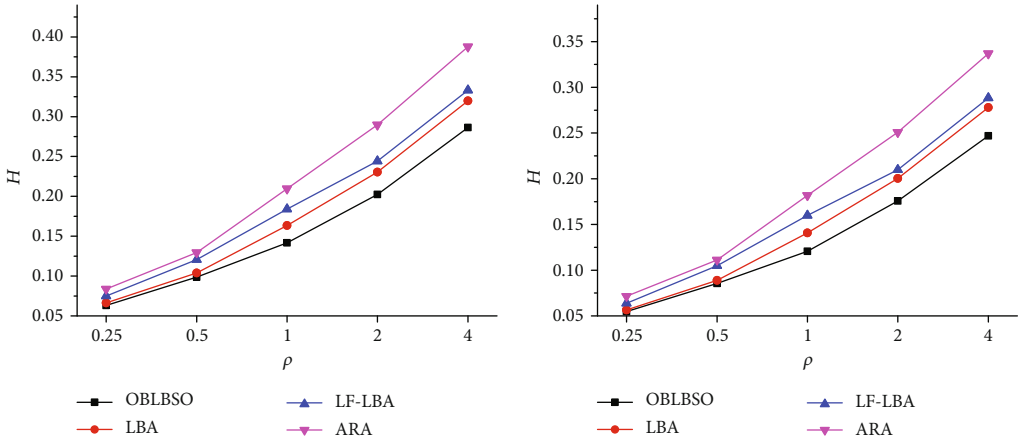
7.1. Parameters Setting

7.1.1. Network Parameters. There are $N_{vnf} = 5$ kinds of VNFs, and each data center can provide 2-5 types of VNFs. Each VNF-sc requires at least one VNF, and the required frequency slots meet uniform distribution [5, 10]. In addition, in [0.5, 1.5], the required frequency-to-slot ratio satisfies a uniform distribution after implementing the corresponding VNFs. Each fiber can accommodate 1000 frequency slots, that is, $N_F = 1000$. In the proposed improved brain storm optimization algorithm, the following parameters are chosen: population size $P_s = 100$, maximum iterations $G_{\max} = 30{,}000$, $\tau_1 = \tau_2 = \tau_3 = 2$.

7.2. Experimental Results. And compared with several other algorithms, the algorithm proposed in the literature [33] (LBA for short) is used for improvement. The second is the LF-LBA algorithm, including minimum priority strategy and LBA algorithm. In addition, we also compared OBLBSO

1 N individuals are randomly generated;
2 The individuals were clustered in the decision space and divided into G clusters;
3 Evaluate individual fitness value;
4 The individuals in each cluster were sorted and the best individuals in each cluster were recorded as the center of the cluster;
5 A random number between 0 and 1 is generated randomly. If the random number is less than the preset probability of p_1, an individual will be generated randomly to replace a cluster center;
6 Randomly generate a number p_3 between 0 and 1;
7 **if** p_3 is less than the default probability of p_2 **then**
8 Random into a random number between 0 and 1, and randomly choose a cluster;
9 If the random number is less than p_4, the center of the cluster is selected and a new individual is generated through Gaussian variation;
10 Otherwise, other individuals in the cluster are selected and new individuals are generated by Gaussian variation;
11 **else**
12 Randomly generate a number between 0 and 1;
13 If the random number is less than p_5, then a new individual is generated based on the center of the two clusters through Gaussian variation. Otherwise, two individuals selected at random based on two clusters will be generated by Gaussian variation;
14 **end**
15 The newly generated individuals were compared with the existing ones, and the well-preserved individuals were taken as the next generation of new individuals;
16 If a preset maximum number of iterations is reached, stop, or skip to Step 2.

Algorithm 2: VNF-SC mapping algorithm-based OBLBSO.



(a) $N_D = 1/3N_V$ (b) $N_D = 2/3N_V$

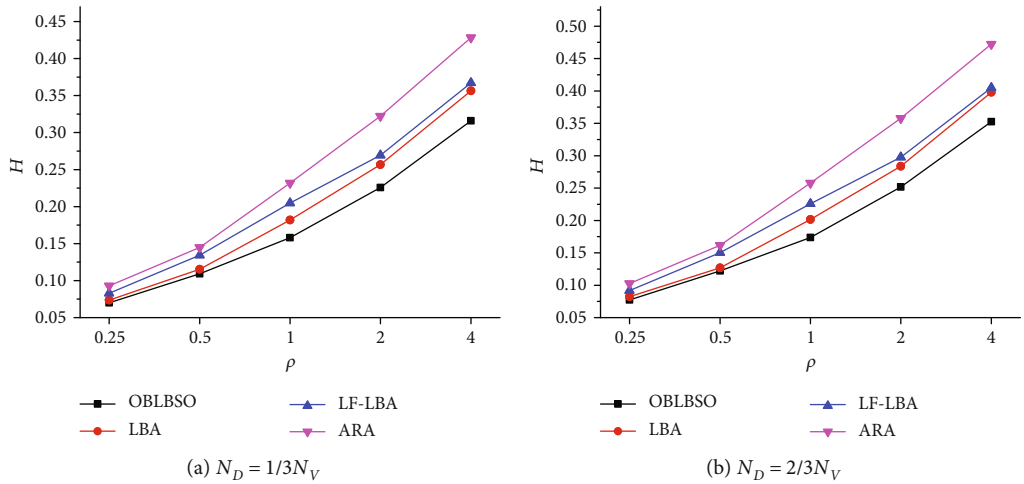
Figure 1: Experimental results when $\alpha_1 = 1$.

with the ARA (artificial raindrop algorithm, ARA) proposed in the literature [34].

For the sake of verifying the performance of the model and algorithm, two experimental scenarios were carried out. In the first scenario, we fixed the number of target nodes as $N_D = N_V/3$ and $N_D = 2N_V/3$, i.e., $N_d = N_D/N_V = 1/3$ and $N_d = N_D/N_V = 2/3$. The number of destination nodes generated in $[N_V/6, N_V/3]$ and $[N_V/3, 2N_V/3]$ randomly. Figure 1 shows the experimental results when $\alpha = 1$. Experimental results when $\beta = 1$ is shown in Figure 2. Figure 3 shows the experimental results when $\gamma = 1$. Experimental results when $\alpha = \beta = \gamma = 1/3$ is shown in Figure 4. Number of connection requests are set as $N_R = \rho N_V(N_V - 1)$, and ρ =0.25, 0.5, 1, 2, and 4, respectively.

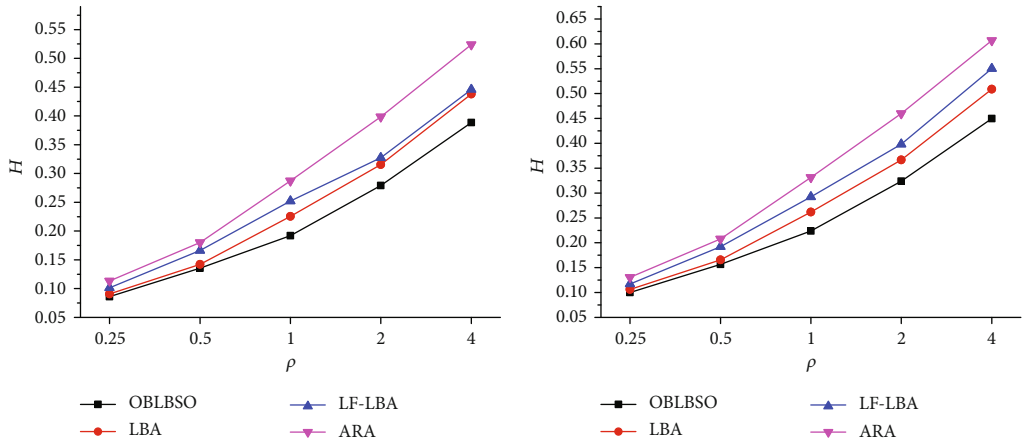
In the second scene, we fixed α_1, α_2 and α_3 to $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$. Figure 5, respectively, shows $\rho = 0.25$s, $\rho = 0.5$, $\rho =$ 1, $\rho = 2$, $\rho = 4$, and $\rho = 8$. The result obtained in the NSFNST topology at the time. In each experiment, set the number of connection requests to $N_D = \theta N_V$ and select $\theta =$ to be 0.2, 0.4, 0.6, 0.8, and 1, respectively.

7.3. Experimental Analysis. The experimental results obtained under NSFNET topology are shown in the figure. When α_1, α_2, and α_3 are selected as 1, 0, 0 1, therefore, the objective function is to minimize the maximum use frequency slot index (MIUFS). It can be seen from the experimental that the OBLBSO achieves better results than the other algorithm. Generally speaking, the minimum priority strategy can reduce the maximum index of the frequency slot used. However, VNF dependencies can disable it. Therefore, in some cases, LBA can achieve a better solution than LF-LBA, and in other cases, LF-LBA is a better solution



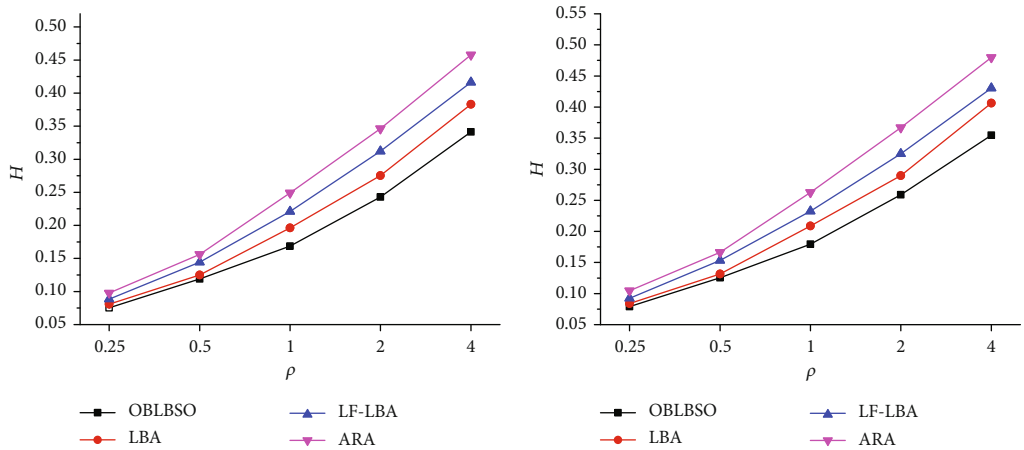
(a) $N_D = 1/3N_V$ (b) $N_D = 2/3N_V$

Figure 2: Experimental results when $\alpha_2 = 1$.



(a) $N_D = 1/3N_V$ (b) $N_D = 2/3N_V$

Figure 3: Experimental results when $\alpha_3 = 1$.



(a) $N_D = 1/3N_V$ (b) $N_D = 2/3N_V$

Figure 4: Experimental results when $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$.

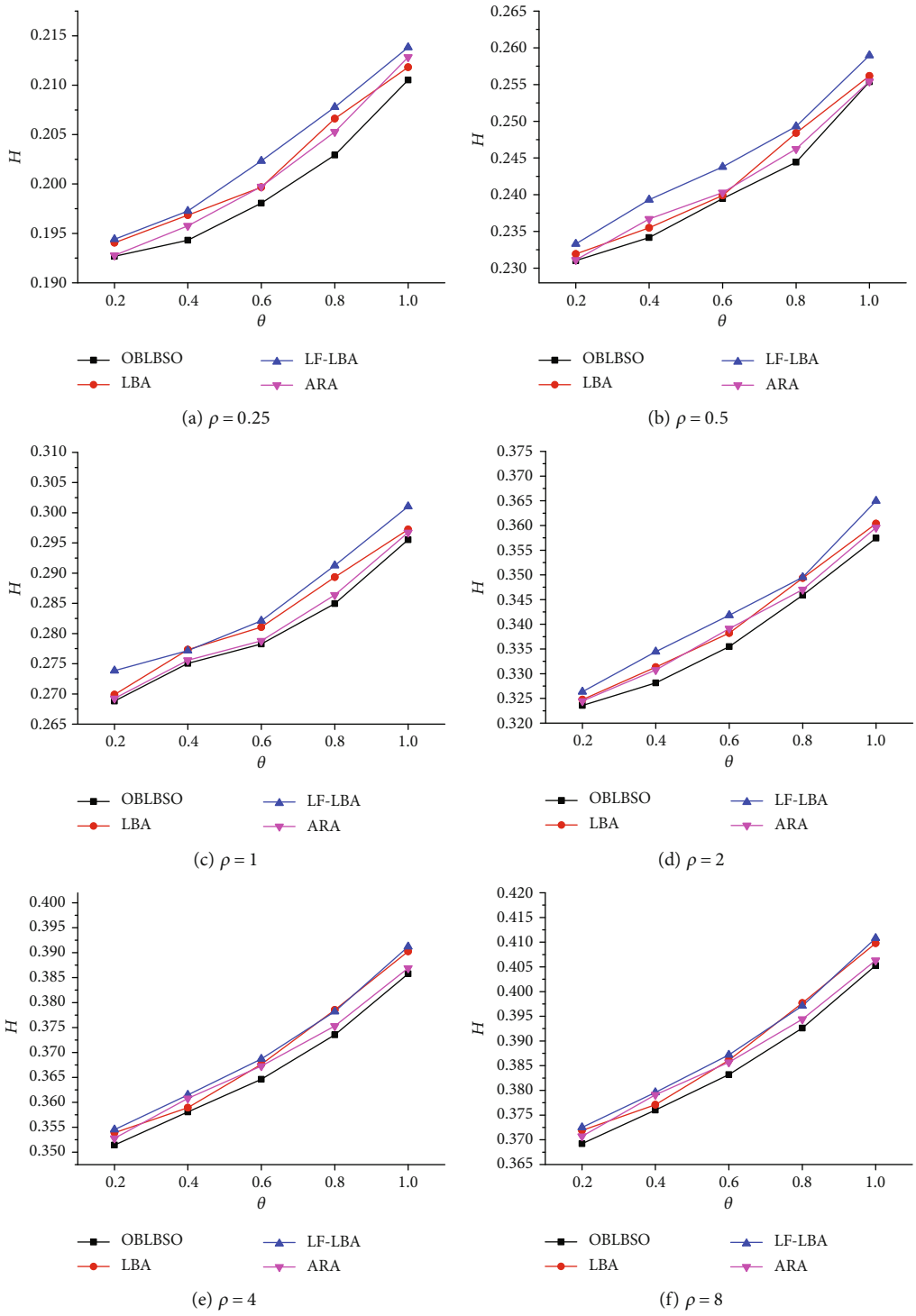


Figure 5: Experimental results obtained when $\rho = 0.25,0.5,1,2,4,8$.

than LBA. OBLBSO can find the optimal routing and VNFs deployment plan for all VNF-SCs. Therefore, in the three algorithms, OBLBSO can get the optimal solution. When the number of connection requests is $0.25N_V(N_V-1)$, the MIUFS obtained by OBLBSO is 3.7%-4.5% less than the MIUFS obtained by the other algorithms. When the number of connection requests is $2N_V(N_V-1)$, the MIUFS obtained by OBLBSO is 7.9%-9.0% less than the MIUFS obtained by the other algorithms, respectively. In other words, as the number of connection requests increases, OBLBSO can obtain a smaller total power consumption and save more power than other algorithms.

When α, β, and γ are selected as 0,1,0, the experimental results are shown in Figure 2. Similar to the experimental results in Figure 1, we can also see that the OBLBSO achieves better results than the other algorithms. In addition, based on

the experimental results, we cannot distinguish between LBA and LF-LBA.

Figure 3 shows the experimental results when α_1, α_2, and α_3 are selected as 1, 0, and 0. Figure 4 shows the experimental results at $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$. It can be seen from the experimental results that the OBLBSO can obtain a better solution than the two other algorithms.

It can be seen from the experimental results that using OBLBSO can get better results than ARA. On this basis, the individual location update strategy has been improved. This algorithm, like the particle swarm algorithm and the DE algorithm, uses the location of other individuals and their past location information, thereby enhancing the search capability and improving the convergence speed.

8. Conclusion

The deployment of VNFs of VNF-SC in flexible optical networks between data centers is studied. In an elastic optical network between data centers, each data center can only provide specific VNFs, and system resources are limited. A mixed integer linear programming model is established, and an improved brainstorming optimization algorithm (OBLBSO) is proposed to solve the model. A simulation experiment was carried out on a widely used network topology.

Data Availability

All the data can be found in the manuscript.

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 (No. 62006205, 62002307), Innovation Team Support Plan of University Science and Technology of Henan Province (No. 19IRTSTHN014), Foundation of Henan Educational Committee under Contract (No. 21A520039), and Nanhu Scholars Program for Young Scholars of XYNU, Youth Sustentation Fund of Xinyang Normal University (No.2019-QN-040).

References

[1] D. Bhamare, R. Jain, M. Samaka, and A. Erbad, "A survey on service function chaining," *Journal of Network and Computer Applications*, vol. 75, pp. 138–155, 2016.

[2] B. Blanco, I. Taboada, J. Oscar Fajardo, and F. Liberal, "A robust optimization based energy-aware virtual network function placement proposal for small cell 5g networks with mobile edge computing capabilities," *Mobile Information Systems*, vol. 2017, no. 4, pp. 1–14, 2017.

[3] C. Galdamez, R. Pamula, and Z. Ye, "On efficient virtual network function chaining in nfv-based telecommunications networks," *Cluster Computing*, vol. 22, no. 3, pp. 693–703, 2019.

[4] J. Masahiko, "Virtualization in optical networks: From elastic networking level to sliceable equipment level," in *The 10th International Conference on Optical Internet (COIN2012)*, pp. 61-62, Yokohama, Kanagawa, Japan, May 2012.

[5] B. Chen, J. Zhang, W. Xie, J. P. Jue, Y. Zhao, and G. Shen, "Cost-effective survivable virtual optical network mapping in flexible bandwidth optical networks," *Journal of Lightwave Technology*, vol. 34, no. 10, pp. 2398–2412, 2016.

[6] X. Xingsi and L. Jiawei, "A compact brain storm algorithm for matching ontologies," *IEEE Access*, vol. 8, pp. 43898–43907, 2020.

[7] B. Guo, C. Qiao, J. Wang et al., "Survivable virtual network design and embedding to survive a facility node failure," *Journal of Lightwave Technology*, vol. 32, no. 3, pp. 483–493, 2013.

[8] R. Lu and X. Nan, "Survivable multipath routing and spectrum allocation in OFDM-based flexible optical networks," *Journal of Optical Communication and Network*, vol. 5, no. 3, pp. 172–182, 2013.

[9] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, p. 2056, 2020.

[10] H. Kim, "Performance evaluation of revised virtual resources allocation scheme in network function virtualization (nfv) networks," *Cluster Computing*, vol. 22, no. S1, pp. 2331–2339, 2019.

[11] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and H. Iztok, "Mobility-aware caching and computation offloading in 5g ultra-dense cellular networks," *Sensors*, vol. 16, no. 7, p. 974, 2016.

[12] L. Fang, X. Zhang, K. Sood, Y. Wang, and S. Yu, "Reliability-aware virtual network function placement in carrier networks," *Journal of Network & Computer Applications*, vol. 154, p. 102536, 2020.

[13] K. Odagiri, S. Shimizu, and N. Ishii, "Establishment of virtual policy based network management scheme by load experiments in virtual environment," *International Journal of Computer Networks & Communications*, vol. 8, no. 3, pp. 181–194, 2016.

[14] R. J. Pfitscher, A. S. Jacobs, L. Zembruzki et al., "Guiltiness: a practical approach for quantifying virtual network functions performance," *Computer Networks*, vol. 161, pp. 14–31, 2019.

[15] S. I. Kuribayashi, "Virtual routing function deployment in nfv based networks under network delay constraints," *International Journal of Computer Networks & Communications*, vol. 10, no. 1, pp. 35–44, 2018.

[16] A. D. Domenico, Y. F. Liu, and W. Yu, "Optimal virtual network function deployment for 5g network slicing in a hybrid cloud infrastructure," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7942–7956, 2020.

[17] R. Yu, G. Xue, V. T. Kilari, and X. Zhang, "Network function virtualization in the multi-tenant cloud," *IEEE Network*, vol. 29, no. 3, pp. 42–47, 2015.

[18] V. Malathi, S. Takehiro, B. Molly, R. Reza, O. Satoru, and Y. Naoaki, "Network function virtualization: a survey," *IEICE Transactions on Communications*, vol. E100.B, no. 11, pp. 1978–1991, 2017.

[19] K. Rakesh, M. Manoj, and A. K. Sarje, "A proactive load-aware gateway discovery in ad hoc networks for internet connectivity," *International Journal of Computer Networks & Communications*, vol. 2, no. 5, pp. 275–282, 2010.

[20] Q. Zheng, R. Li, X. Li et al., "Virtual machine consolidated placement based on multi-objective biogeography- based

optimization," *Future Generation Computer Systems*, vol. 54, pp. 95–122, 2016.

[21] D. Bhamare, A. Erbad, R. Jain, M. Zolanvari, and M. Samaka, "Efficient virtual network function placement strategies for cloud radio access networks," *Computer Communications*, vol. 127, pp. 50–60, 2018.

[22] V. Eramo, E. Miucci, M. Ammar, and F. G. Lavacca, "An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2008–2025, 2017.

[23] T. Wen, H. Yu, G. Sun, and L. Liu, "Network function consolidation in service function chaining orchestration," in *IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, 2016.

[24] A. Hameed, A. Khoshkbarforoushha, R. Ranjan et al., "A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems," *Computing*, vol. 98, no. 7, pp. 751–774, 2016.

[25] L. Qu, M. Khabbaz, and C. Assi, "Reliability-aware service chaining in carrier-grade softwarized networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 558–573, 2018.

[26] A. N. Toosi, J. Son, Q. Chi, and R. Buyya, "Elasticsfc: auto-scaling techniques for elastic service function chaining in network functions virtualization-based clouds," *Journal of Systems and Software*, vol. 152, pp. 108–119, 2019.

[27] L. Tang, G. Zhao, C. Wang, P. Zhao, and Q. Chen, "Queue-aware reliable embedding algorithm for 5g network slicing," *Computer Networks*, vol. 146, pp. 138–150, 2018.

[28] H. Xuan, Y. Wang, Z. Xu, S. Hao, and X. Wang, "Virtual optical network mapping and core allocation in elastic optical networks using multi-core fibers," *Optics Communications*, vol. 402, pp. 26–35, 2017.

[29] H. Xuan, S. Wei, Y. Feng, H. Guo, and Y. Li, "A new bi-level mathematical model and algorithm for vons mapping problem," *IEEE Access*, vol. 8, pp. 101797–101811, 2020.

[30] S. P. Singh and S. C. Sharma, "Implementation of a PSO based improved localization algorithm for wireless sensor networks," *IETE Journal of Research*, vol. 65, pp. 502–514, 2018.

[31] H. R. Tizhoosh, "Opposition-based learning: a new scheme for machine intelligence," in *International Conference on International Conference on Computational Intelligence for Modelling, Control & Automation*, pp. 695–701, Vienna, Austria, Nov 2005.

[32] S. Hui, D. Zhi-Cheng, Z. Jia, W. Hui, and X. Hai-Hua, "Hybrid mean center opposition-based learning particle swarm optimization," *Acta Electronica Sinica*, vol. 47, no. 9, pp. 1809–1818, 2019.

[33] W. Fang, M. Zeng, X. Liu, W. Lu, and Z. Zhu, "Joint spectrum and it resource allocation for efficient vnf service chaining in inter-datacenter elastic optical networks," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1539–1542, 2016.

[34] Q. Jiang, L. Wang, Y. Lin, X. Hei, and X. Lu, "An efficient multi-objective artificial raindrop algorithm and its application to dynamic optimization problems in chemical processes," *Applied Soft Computing*, vol. 58, no. 5, pp. 354–377, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8842508, 11 pages
https://doi.org/10.1155/2021/8842508

Research Article

Soil Medium Electromagnetic Scattering Model for the Study of Wireless Underground Sensor Networks

Frank Kataka Banaseka,[1,2] **Hervé Franklin**,[2] **Ferdinand A. Katsriku**,[2] **Jamal-Deen Abdulai**,[2] **Akon Ekpezu**,[2] **and Isaac Wiafe**[2]

[1]*University of Professional Studies, Accra, Department of Information Technology, Ghana*
[2]*University of Ghana, Legon, Department of Computer Science, Ghana*

Correspondence should be addressed to Frank Kataka Banaseka; frank.banaseka@upsamail.edu.gh

Received 9 September 2020; Revised 26 November 2020; Accepted 22 December 2020; Published 4 January 2021

Academic Editor: Pei-Wei Tsai

In recent years, there has been keen interest in the area of Internet of Things connected underground, and with this is the need to fully understand and characterize their operating environment. In this paper, a model, based on the Peplinski principle, for the propagation of waves in soils that takes into account losses attributable to the presence of local inhomogeneity is proposed. In the work, it is assumed that the inhomogeneities are obstacles such as stones or pebbles, of moderate size, all identical and randomly distributed in space. A new wave number is obtained through a combination of the multiple scattering theory and the Peplinski principle. Since the latter principle considers the propagation in a homogeneous medium (without obstacles), the wave number it provides is inserted into the one resulting from the former, the multiple scattering theory. The effective wave number thus obtained is compared numerically with that of Peplinski alone on the one hand and with that of multiple scattering alone on the other hand. The phase velocity and the loss tangent are analyzed against the particle concentration at the low-frequency Rayleigh limit condition ($ka \lesssim 0.1$) and against the frequency at two particle concentrations ($\boldsymbol{c} = \mathbf{0.2}$ and $\boldsymbol{c} = \mathbf{0.4}$), two particle radii ($\boldsymbol{a} = \mathbf{0.55}$ cm and $\boldsymbol{a} = \mathbf{1.10}$ cm), and 5% and 50% volumetric water content of the soil. Path losses are also compared to each other to examine the effects on transmission of soil containing obstacles. The results obtained suggest that the proposed model has better accuracy in estimating the wave number than previously used schemes.

1. Introduction

Wireless underground sensor networks (WUSNs) are an emerging area that has gained the attention of many researchers. This is because WUSNs open up new possibilities for underground monitoring and communication; also, they will find application in agriculture, which is key to the developmental agenda of many emerging nations. For instance, a WUSN path loss model based on an accurate prediction of the complex dielectric constant (CDC) for precision agriculture is proposed and called WUSN-PLM [1]. These WUSNs are now being interconnected to form the "Internet of Underground Things (IoUT)" to depict the internet of devices connected underground. IoUT encompasses devices buried in soil or placed in open bounded spaces and are interconnected to facilitate sending of information out of agricultural fields and other underground environments to decision-making and control centres. In the same context, an analytical survey was performed on the current and potential application of the Internet of Things in arable farming, state-of-the-art technologies deployed, challenges of mobile devices in spatial data collection, highly varying environments, and task diversity, compared to other agricultural systems [2, 3].

The physical phenomenon underpinning the operation of IoUT is the propagation of electromagnetic waves in a soil medium. The characteristics of the operating environment will have a great impact on the performance of the network. Obstacles such as stones cause waves to be refracted or scattered in an underground environment. In addition, an increase in communication range and volumetric water content of soil due to irrigation or rain will lead to high signal

path loss. Additionally, signal propagation characteristics in soil are dependent on the soil type and properties. Typically, a two-stage model is proposed based on the field characteristics of the antenna and considers four sources of path loss. The two-stage model has a different coefficient, which depends on the soil types in the near-field and far-field regions [4, 5].

To effectively characterize the propagation environment, accurate and robust models are required. Different models have been proposed in the literature for the study of electromagnetic wave propagation in various environments [6–14]. In particular, for the characterization of electromagnetic wave (EM) propagation in soil, a number of different models have been proposed in the literature [6, 11, 12, 14]; however, models that take into consideration factors such as multipath propagation, volumetric water content of soil, and burial depth are dominant [6, 11]. Furthermore, some of these models analyze the bit error rate for communication performance based on some modulation schemes and soil properties.

As an alternative to EM wave propagation in soil, Sun and Akyildiz [11] have proposed the magnetic induction technique. The magnetic induction technique is a promising communication technique for analyzing propagation in soil [10, 11]. EM waves and the Friis equations [12] were used to analyze the channel model taking into consideration the direct, reflected, and lateral waves, multipath, soil composition, and water content. It was shown that the direct, reflected, and lateral waves are major contributors to signal attenuation in the soil environment [11]. In [14], a segmentation approach is used to sense soil moisture where the radio field is used as a sensor. Based on the Peplinski principle [15], the path loss for different volumetric water content levels at three different frequencies was calculated. In comparison with some related research proposed recently as shown in Table 1, this work seeks to analyze the transmission of electromagnetic waves in a soil medium taking into account the presence of obstacles that cause multiple scattering. A new wave number model is proposed with the combination of the Peplinski principle and multiple scattering in the soil medium. The new wave number is used in the computation of the path loss.

To the best of our knowledge, path loss expressions that consider explicitly scattering and in particular multiple scattering are yet to be reported. In this paper, based on the results presented in [9, 14] where a relation for the path loss is derived, we consider the problem of EM wave propagation [16–18] in a dense medium with scattering properties. In the present work, a model of the effective wave number is presented that accounts for absorption due to permittivity and multiple scattering occurring in soil because of the presence of buried obstacles such as stones, rocks, or pebbles. To achieve this, it is assumed that the medium which typically is polydispersed contains identical objects of similar size. This assumption allows the change in path loss to be readily estimated. The propagation constants derived from the effective wave number obtained are used for the calculation of the path loss. The results are compared with those previously reported in the literature.

In addition, a parametric study is also performed that shows the effects of the concentration of obstacles or the volumetric water content on signal attenuation. The phase velocity and the loss tangent are analyzed against the particle concentration at the low-frequency Rayleigh limit condition and against the frequency at two particle concentrations: 0.2 and 0.4, two particle radii: 0.55 cm and 1.10 cm, and 5% and 50% volumetric water content of the soil. The analysis is performed considering only the Peplinski principle on the one hand and the Peplinski principle with multiple scattering on the other. Results obtained indicate that the approach proposed in this study could provide significantly better results than previously obtained and lead to a better characterization of WUSN.

2. Comparison with Some Related Research Proposed Recently

In [14], Liedmann et al. presented the path loss of an average topsoil for different distances and typical IoT frequencies at two different VWCs, 5% and 50%, respectively. The higher the operating frequency, the higher the influence of rising VWC. Meanwhile, in this work, path loss is modeled based on absorption due to permittivity and multiple scattering from obstacles in soil. It is then analyzed against distance at the same frequencies of 433 MHz and 868 MHz and at the same VWC proportions. Path loss analysis in both works shows almost the same trends.

In [1], four sources of path losses are analyzed based on a proposed two-stage model with field characteristics of the antenna. The two-stage model has a different coefficient, which depends on the soil types in the near-field and far-field regions. Path losses against transmission distance are compared on dry soil for sandy clay#1 and for sandy clay#2. Path losses against sensor burial depth for underground-to-above the ground and above the ground-to-underground channel models are also compared for 5 m, 10 m, 15 m, and 20 m transmission ranges. Results in this work show the same trend of path loss.

At the same operating frequency of 433 MHz in [4], the comparison is made between the measured path loss and that of Friis, Fresnel, and the proposed propagation models tested within clayey silt, dry sand, and wet sand media. The results showed the same trend of increasing growth of path loss as the transmission distance increases to 0.5 m, 1 m, and beyond. Meanwhile, in this work, a comparison has been established for a revised path loss based on only the Peplinski principle and on Peplinski combined with multiple scattering for two operating frequencies 868 MHz and 433 MHz. The transmission distance considered in our study is up to 5 m.

3. System Architecture and IoUT Application in Agriculture

In agriculture, IoUT is envisaged to provide total field autonomy and enable more efficient food production solutions through not only in situ monitoring and self-reporting capabilities (soil moisture, salinity, temperature, etc.) but also the interconnection of existing field machinery like irrigation

Table 1: Comparison of some related research proposed recently with our work.

Reference	Related research proposed recently	Our work
[1]	A WUSN path loss model for precision agriculture called WUSN-PLM is proposed. The proposed model is based on an accurate prediction of the complex dielectric constant (CDC).	A new wave number model is proposed using the combination of the Peplinski principle and multiple scattering in a soil medium. The new wave number is used in the computation of the path loss.
[2]	Underground environment-aware MIMO is developed using transmit and receive beamforming.	Signal transmission with single input single output (SISO) between a transmitter and a receiver.
[3]	Analytical survey of the current and potential application of the Internet of Things in arable farming, state-of-the-art technologies deployed, challenges of mobile devices in spatial data collection, highly varying environments, and task diversity, compared to other agricultural systems.	Internet of Underground Things (IoUT) application in precision agriculture, envisaged to provide total field autonomy and enable more efficient food production solutions through not only in situ monitoring and self-reporting capabilities but also the interconnection of existing field machinery like irrigation systems, harvesters, and seeders.
[4]	A two-stage model is proposed based on the field characteristics of an antenna and considers four sources of path loss. The two-stage model has a different coefficient, which depends on the soil types in the near-field and far-field regions.	Path loss is modeled based on absorption and multiple scattering from obstacles in soil. The path loss is then analyzed against distance at two typical IoT frequencies of 433 MHz and 868 MHz. We also showed the effect of VWC on the path loss for two proportions, 5% and 50%.

systems, harvesters, and seeders [19] as shown in Figure 1. Real-time decision-making takes place in IoUT at a monitoring centre in the cloud, which receives information required from sensors, underground, and other field devices. It is worth noting that in this architecture communication may take place through the soil medium between the underground devices or may take place through air and plant foliage devices attached to the plant body. IoUT applications have challenging requirements that include impairments attributable to wireless communication through soil [20–23] and machinery operating in remote crop fields and exposure to the elements. Progress in IoUT research will significantly impact and benefit many application areas such as plant monitoring and control, border patrol, underground mines and tunnels, and pipeline assessment [6, 20, 22]. A model of IoUT architecture is illustrated in Figure 1. Based on the required functionalities, the architecture may include components such as underground objects, base stations, mobile sinks, and cloud services. Below, we explain the properties or characteristics of the various components.

Underground objects (UOs) are made up of objects which are wholly or partially embedded within the subterranean environment and are equipped with wireless communication and sensing modules. They are constructed to be rugged enough to protect them against farming equipment, wild rodents, and extreme weather conditions. Their onboard sensors usually include a wide range of soil-related or weather-related sensing devices that can monitor phenomena such as soil temperature, chemical properties, and moisture. The communication schemes employed include Bluetooth, ZigBee, NFC, Wi-Fi, Sigfox, LoRa, LoRaWAN, Satellite, and cellular [24].

Base stations (BSs) have high processing power and communication facilities installed in permanent structures like weather stations or buildings and are used as gateways to transfer the data collected from wireless nodes to the cloud (the monitoring centre). While it is not essential to have a base station in a wireless communications framework, they can be key to the operation of a wireless network and in extending its communication range.

In wireless sensor network applications, data transfer can account for up to 70% of the energy cost. This has led to the use of mobile sinks for data collection. Mobile sinks are vehicles (tractors, irrigation systems, harvesters, seeders, and unmanned aerial vehicle drones) that move periodically or as required, around the defined fields where the sensors are deployed to collect data from the sensor nodes.

Cloud services are used for real-time processing of the field data as well as permanent storage of the collected data. The cloud service, as the decision-making centre of the system, helps in integrating collected data with existing databases. This also serves as the platform through which services and information are made available to users and stakeholders.

Within the IoUT architecture, a number of different communication links may be identified. This includes (a) underground-to-underground link in which the entire communication is confined underground, such a scheme might be used when the sensors need to exchange information or when a sensor needs to route data through a neighbouring sensor; (b) underground to above ground: this part of the communication is underground and the other part is above ground with the transmitter being buried underground and the receiver is above. This scheme is mainly used to send data collected by the sensors to control centres; (c) above ground to underground: here, the transmitter is above ground, and the receiver may be the sensor device underground. The scheme is mainly used for sending control signals; (d) above-ground surface communication. In this deployment, the sensors (transmitter and receiver) are partially buried in the ground as such the communication link is above ground but within the immediate vicinity of the soil, the wave will therefore propagate just above the surface of the ground. Arguably, existing over-the-air (OTA) wireless communication protocols face significant challenges in underground environments because they were originally not designed for

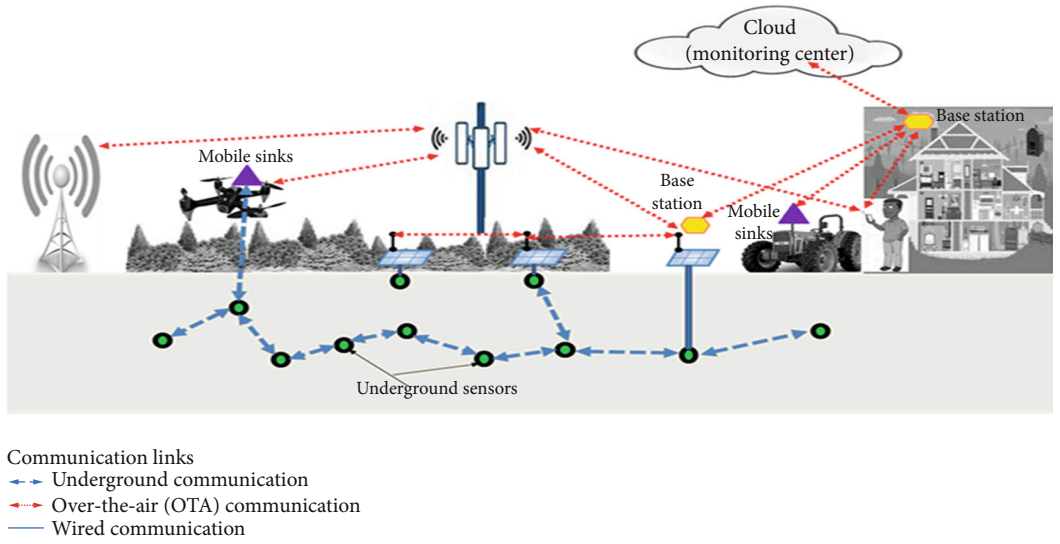


Figure 1: Architecture of IoUT.

these conditions. The electromagnetic wave attenuation in soil is much higher than that in air. This has a limiting effect on link quality. In above-ground-to-underground communication, it is critical to give due consideration to the important effects of reflection and refraction due to the ground surface.

4. Absorption due to Permittivity

In the following, a uniform random distribution of dielectric spheres with radii a (also referred to as particles) of relative permittivity ε_p embedded in a background medium (soil) of permittivity ε is considered. Let n_o be the number of spheres per unit volume and α the polarizability of each sphere. The polarization $\mathbf{P}$ (dipole moment per unit volume) is given by

$$\mathbf{P} = \frac{n_o \alpha}{1 - n_o \alpha / 3\varepsilon} \mathbf{E}, \tag{1}$$

where $\mathbf{E}$ is the electric field inside the medium. Substituting (1) into the electric flux density $\mathbf{D} = \varepsilon \mathbf{E} + \mathbf{P}$ yields $\mathbf{D} = \varepsilon_{\text{eff}} \mathbf{E}$, where

$$\varepsilon_{eff} = \varepsilon \left[\frac{1 + 2n_o \alpha / 3\varepsilon}{1 - n_o \alpha / 3\varepsilon} \right] \tag{2}$$

are Clausius-Mossotti's formula of the effective permittivity. The formula for the polarizability is [3]

$$\alpha = 3\varepsilon v_o \frac{\varepsilon_p - 1}{\varepsilon_p + 2}, \tag{3}$$

where $v_0 = 4\pi a^3/3$ is the volume of the sphere. Substituting α from (3), the relation for the effective permittivity yields Maxwell-Garnett's mixing formula:

$$\varepsilon_{eff} = \varepsilon \left(\frac{1 + 2cy}{1 - cy} \right), \tag{4}$$

where $c = n_o v_o$ is the fractional volume occupied by the particles and y is given by

$$y = \frac{\varepsilon_p - 1}{\varepsilon_p + 2}. \tag{5}$$

The effective wave number K of the composite medium is given by $K = \omega \sqrt{\mu \varepsilon_{eff}}$ or

$$K^2 = k^2 \frac{1 + 2cy}{1 - cy} = k^2 \left(1 + \frac{3cy}{1 - cy} \right), \tag{6}$$

where $k = \omega \sqrt{\mu \varepsilon}$ is the wave number for the background medium free of spheres and having a permittivity. The existence of an imaginary part for K reflects the presence of the absorption phenomenon in the medium during the propagation. In (6), the permittivity can be derived from the Peplinski principle which is discussed later.

5. Wave Attenuation due to Multiple Scattering in Soil: Quasicrystalline Approximation (QCA) in Dense Media

Multiple scattering phenomena occur in soil in the presence of inhomogeneity or spheres which are randomly (or not) distributed. The presence of such inhomogeneities can have a significant effect on the propagation of the electromagnetic wave within the soil medium. It can be observed that (6) does not take into consideration the multiple scattering of the waves by the spheres. It seems appropriate to add a term taking into account this multiple scattering.

To address the difficult problem of wave propagation in soils and in particular the problem of multiple scattering, it is helpful to simplify the problem in the first instance. To do this, the scattering simulations can be performed on a test volume containing a large number of spheres but forming at the same time a small part (the representative elementary

volume) of the whole system. The soils are normally polydisperse media since they contain scatterers of various shapes, sizes, and materials. In this work, for simplicity, it is assumed that all the scatterers are identical (monodispersity), the reason being that, by using the simplest equations of multiple scattering, we would be able to estimate the change in path loss. Figure 2 shows a plane electromagnetic wave incident onto a half-space of identical dielectric spheres. Such an incident wave will be subject to multiple scattering in the soil medium. To solve this problem of multiple scattering, the quasicrystalline approximation (QCA) and the T-matrix formalism were used by Tsang et al. [17].

In QCA, statistical configurational averaging using conditional averaging on positions is performed. The details of the calculation techniques leading to the formula of the effective wave number are out of scope for this study. Rather, our interest is on the formulas of the wave number that accounts for attenuation due to scattering.

Let T_n denote the scattering coefficient for a sphere in mode n. Then, at the low-frequency Rayleigh limit ($ka \lesssim 0.1$), most of the contribution is attributable to $T_1(ka)$ (the mode $n = 1$ corresponds to the electric dipole). In the context of the QCA, it follows that the effective wave number [16] is given by

$$K^2 = k^2\left\{1 + \frac{3cy}{1-cy}\left[1 + i2ck^3a^3\frac{y}{1-cy}(1 + 4\pi n_0 J)\right]\right\}, \quad (7)$$

where

$$J = \int_0^\infty r^2[g(r) - 1]dr. \quad (8)$$

In (8), g is the pair distribution function of one sphere position given the position of the other. This function can assume different functional forms, one of the most useful being the Percus-Yevick pair distribution function for hard spheres [19]. The pair distribution function depends on the scatterer size a and the fractional volume c of scatterers within the volume. Different forms of the Percus-Yevick pair distribution function have been discussed in [25] for various values of the fractional volume. In this study, the case of $c = 0.2$ and $c = 0.4$ is considered. The Percus-Yevick function oscillates in the range of $r/2a$ with the degree of oscillation ranging from 1 to 3, depending on the value of c, the fractional volume. The Percus-Yevick approximation for short-ranged hard-sphere pair distribution function is considered [17]:

$$n_0\int_0^\infty r^2[g(r) - 1]dr = \frac{(1-c)^4}{(1+2c)^2} - 1. \quad (9)$$

A simplified formula of the effective wave number may be obtained as follows [18]:

$$K^2 = k^2\left\{1 + \frac{3cy}{1-cy}\left[1 + i\frac{2k^3a^3}{3}\frac{(1-c)^4}{(1+2c)^2}\frac{y}{1-cy}\right]\right\} \quad (10)$$

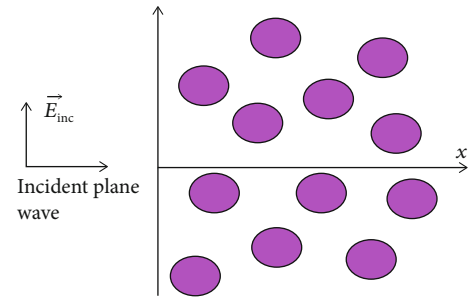


Figure 2: Plane electromagnetic wave normally incident on a half-space of spherical dielectric scatterers of radius a.

(for values of c ranging from 0.2to0.4). For $ka \lesssim 0.1$ and at the operating frequency of 433 MHz (which is within the 0.3-1.3 GHz range of validity of Peplinski's dielectric mixing formula [(13)]), (10) is valid for soils with a random distribution of spheres having a radius not exceeding 1.1 cm as shown in Table 2. If the frequency is higher, a smaller radius is needed to satisfy the limiting condition $ka \lesssim 0.1$, which in essence, is an approximate condition [26].

Note that lower frequencies are required for adequate communication in the soil medium. However, reducing the operating frequency below 300 MHz will increase the antenna size, which introduces practical challenges during WUSN implementation. Most wireless underground sensor boards such as MICA2 are designed to operate within 300 to 400 MHz range. Ideally, operating frequencies of 300 and 900 MHz are appropriate for preserving small antenna sizes [6]. This ensures that the sensors remain discrete, a property which is particularly useful for security applications. For a sparse medium (very small volume fraction), the Percus-Yevick function tends to the Hole-Correction formula [27], and the QCA in (10) is reduced to the EFA (Effective Field Approximation) represented by the simpler but less accurate formula:

$$K^2 = k^2\left\{1 + 3cy\left[1 + \frac{2}{3}ik^3a^3y\right]\right\}. \quad (11)$$

Figure 3 shows the normalized phase velocity Re (k/K) and the loss tangent 2 Im K/Re K versus the concentration obtained from (10) for three values of ka. The permittivity of the background medium is$\varepsilon = 8.854 \times 10^{-12}(\mathrm{F\,m^{-1}})$, and the relative permittivity of the spheres is $\varepsilon_p = 3.2$. When the concentration increases, the phase velocity decreases monotonically while the loss tangent increases initially until a maximum value is attained when it begins to decrease.

In Figure 4, the permittivity of the background medium is assumed to be $\varepsilon = 8.854 \times 10^{-12}(\mathrm{F\,m^{-1}})$, for spheres of radius $a \approx 0.55$ cm and fractional volume concentrations of $c = 0.2$ and $c = 0.4$. The figure shows the normalized phase velocity Re (k/K) and the loss tangent 2 Im K/Re K given by (10) versus the frequency. From Figure 4, it appears that regardless of the fractional volume concentration, $c = 0.2$ or $c = 0.4$, the normalized phase velocities remain constant. The waves propagate faster in the medium with the lowest

Table 2: Evaluation of the maximum radius for the validity of (19).

Frequency	Condition $ka \lesssim 0.1$	Number of spheres per unit volume $n_0 = c/v_0$
433 MHz	$a \lesssim 1.1$ cm	≈35, 873 (if $c = 0.2$)
868 MHz	$a \lesssim 0.55$ cm	≈286, 860 (if $c = 0.2$)

concentration of spheres (radius $a \approx 0.55$ cm) but have the highest loss tangent.

Figure 5 depicts a plot of the normalized phase velocity and loss tangent against frequency. The permittivity of the background medium remains the same in Figure 4. The radius considered for the spheres is $a \approx 1.10$ cm at concentrations $c = 0.2$ and $c = 0.4$. We can observe that the normalized velocities are practically identical and therefore do not depend on the radius. Secondly, the loss tangents take greater values in the reduced range of 300 MHz-500 MHz. Table 3 provides a summary of the values used in the calculation of the wave number by QCA.

Peplinski et al. [15] reported the development of a semiempirical dielectric model for soils, covering the 0.3-1.3 GHz range. The model provides expressions for the real and imaginary parts of the relative dielectric constant of a soil medium in terms of the soil's textural composition (sand, silt, and clay fractions), the bulk density and volumetric moisture content of the soil, and the dielectric constant of water, the specified microwave frequency, and physical temperature. A comparison of experimental results measured in this study with predictions based on the semiempirical model shows that the model developed underestimates the real part of the dielectric constant for cases where the moisture content of the soil is high.

Assuming a complex-valued permittivity $\varepsilon = \varepsilon' - i\varepsilon''$, the propagation constants (the attenuation constant ς and the phase shift constant ϖ) are given by [2, 22]

$$\gamma = \varsigma + i\varpi, \tag{12}$$

where

$$\varsigma = \mathrm{Re}\,(\gamma) = \omega\sqrt{\frac{\mu\varepsilon'}{2}\left[\sqrt{1+\left(\frac{\varepsilon''}{\varepsilon'}\right)^2}-1\right]}, \tag{13}$$

$$\varpi = \mathrm{Im}\,(\gamma) = \omega\sqrt{\frac{\mu\varepsilon'}{2}\left[\sqrt{1+\left(\frac{\varepsilon''}{\varepsilon'}\right)^2}+1\right]}. \tag{14}$$

According to the Peplinski principle, the relative dielectric properties of soil in the 0.3 to 1.3 GHz band can be estimated as follows:

$$\varepsilon^* = \varepsilon' - i\varepsilon'', \tag{15}$$

where

$$\varepsilon' = 1.15\left[1+\frac{\rho_b}{\rho_s}\left(\varepsilon_s^{\alpha'}-1\right)+m_v^{\beta'}\varepsilon_{f_w}'^{\alpha'}-m_v\right]^{1/\alpha'}-0.68, \tag{16}$$

$$\varepsilon'' = \left[m_v^{\beta''}\varepsilon_{f_w}''^{\alpha'}\right]^{1/\alpha'}. \tag{17}$$

In the above,

$$\varepsilon_{fw}' = \varepsilon_{w\infty} + \frac{\varepsilon_{w0}-\varepsilon_{w\infty}}{1+(2\pi f\tau_w)^2}, \tag{18}$$

$$\varepsilon_{fw}'' = \frac{2\pi f\tau_w(\varepsilon_{w0}-\varepsilon_{w\infty})}{1+(2\pi f\tau_w)^2} + \frac{\sigma_{\mathrm{eff}}}{2\pi\varepsilon_0 f}\frac{(\rho_s-\rho_b)}{\rho_s m_v} \tag{19}$$

represent the real and imaginary parts of the relative dielectric constant of water, with $\varepsilon_{w\infty}$ the high-frequency limit of ε_{fw}', ε_{w0} the static dielectric constant, and f(Hz) the operating frequency [12]. The other quantities appearing in (16)–(19) are given in Table 4.

In Figures 6 and 7, the red curves show results using Peplinski's principle. Blue curves show the results when both Peplinski (background medium) and multiple scattering are considered. The radius considered for the spheres is $a \approx 1.10$ cm, $c = 0.2$ and $c = 0.4$. Figure 6 shows the graphs of the phase velocity and the loss tangent versus the frequency when the permittivity of the background medium is given by (15) with 5% VWC. This background is introduced in (10) to yield the wave number that combines both Peplinski absorption (Maxwell-Garnett) and the multiple scattering. For a fixed frequency, it can be seen that the velocity decreases as the concentration increases and is much smaller than that in Figure 4. For the loss tangent, the blue and red curves decrease and show the same trend. Figure 7 shows a plot of the normalized phase velocity and loss tangent versus frequency with the volumetric water content (VWC) equal to 50%. For the loss tangent, blue and red curves, although showing the same trend, dissociate as the frequency increases. The radius considered for the spheres is the same as that in Figure 6: $a \approx 1.10$ cm, $c = 0.2$ and $c = 0.4$. The comparisons of Figures 6 and 7 show that increasing the VWC lowers the phase velocity and the loss tangent, and this has a great influence on the wave propagation.

6. Path Loss versus Distance

Using the propagation constant in (12) derived from Peplinski's principle, the path loss L_p^{Pe} of an electromagnetic wave propagating in soil can be expressed as follows [28]:

$$L_p^{Pe} = 6.4 + 20\log(d) + 20\log(\varpi) + 8.69\varsigma d, \tag{20}$$

where d is the distance between sender and receiver and $8.69\varsigma d$ is the additional attenuation caused by transmission. Our mixing of Peplinski's principle (12) with the multiple scattering theory given by (10) consists in replacing the

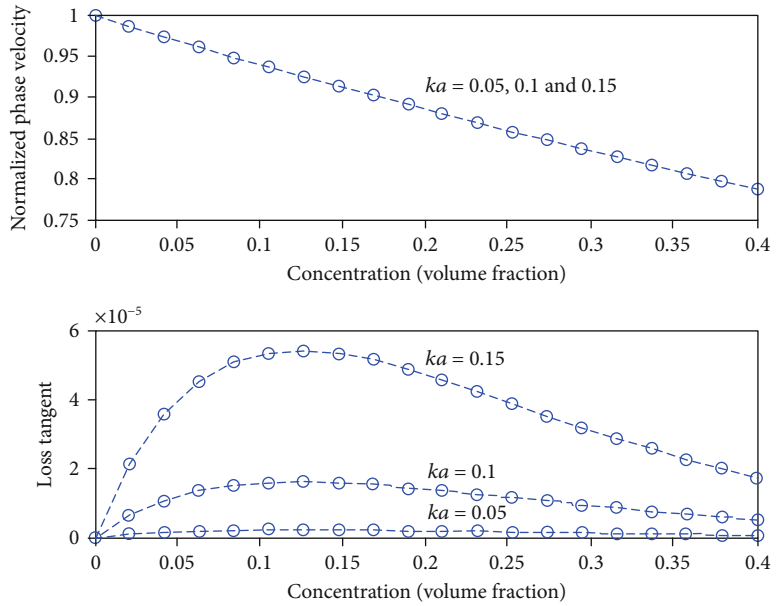


Figure 3: Normalized phase velocity and loss tangent versus concentration obtained from (10).

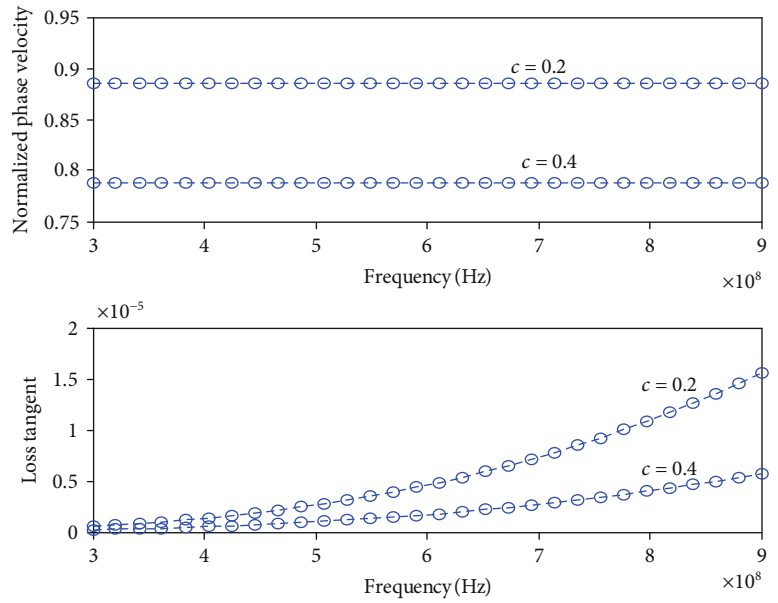


Figure 4: Normalized phase velocity and loss tangent versus frequency from (10) $a \approx 0.55$ cm.

background wave number $k = \omega\sqrt{\mu\varepsilon}$ in which ε is the permittivity in free space by $k = \omega\sqrt{\mu(\varepsilon' - i\varepsilon'')}$, where ε' and ε'' are given by (16) and (17). The resulting wave number is denoted by $K^{Pe+\text{m.s.}} = k_r + ik_i$ (m.s. is used for multiple scattering). By making the correspondences $\varpi \to k_r$ and $\varsigma \to -k_i$, the expression of the path loss in (20), which now includes multiple scattering, becomes

$$L_p^{Pe+\text{m.s.}} = 6.4 + 20\log(d) + 20\log(k_r) - 8.69k_i d. \quad (21)$$

The minus sign in (21) accounts for the fact that when propagating in the direction of increasing x, we have $e^{i\varpi x}e^{\varsigma x}e^{-i\omega t}$ using (12) and $e^{ik_r x}e^{-k_i x}e^{-i\omega t}$ using the wave number from our mixing formula.

In Figure 8, the red curves are obtained from Peplinski's principle and the blue curves from the combination of Peplinski and multiple scattering. The VWC in Figure 8(a) is 5%, and in Figure 8(b), it is 50%. The radius considered for the spheres is once again $a \approx 1.10$ cm, $c = 0.2$. The figure shows the path losses L_p^{Pe} and $L_p^{Pe+\text{m.s.}}$ (in dB) versus the distance d between sender and receiver, at two typical IoT frequencies of 433 MHz and 868 MHz. The data for the considered average soil types are as described in Table 4. As

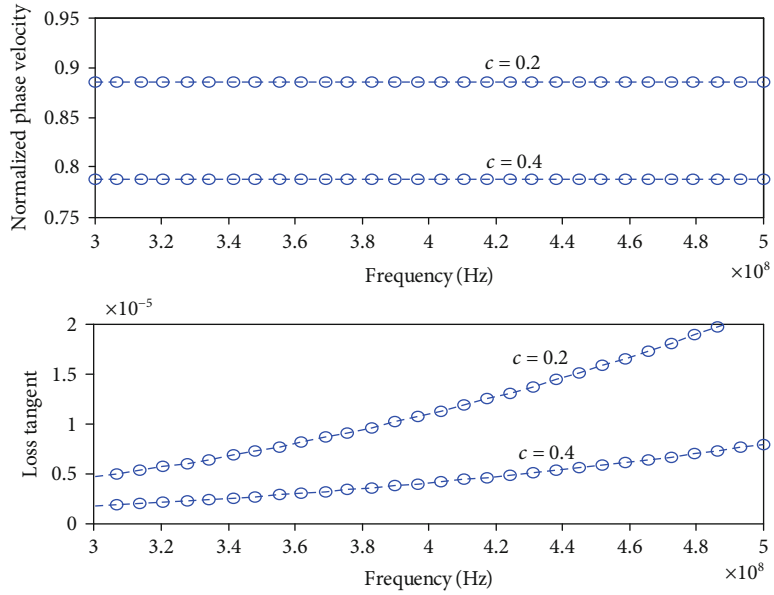


Figure 5: Normalized phase velocity and loss tangent versus frequency from (10) $a \approx 1.10$ cm.

Table 3: The physical parameters used for computing the wave number obtained by QCA.

Symbol	Quantity	Value and units
c	Volume fraction	0.2 or 0.4
ε_p	Relative permittivity in spheres	$3.2 + i0$ (F m^{-1})
ε	Background permittivity	8.854×10^{-12} (F m^{-1})
a	Radius of spheres	0.011 (m) (frequency up to 500 MHz) 0.055 (m) (frequency up to 900 MHz)

Table 4: The physical parameters used for computing propagation constants from Peplinski's principle.

Symbol	Quantity	Value and units
ρ_b	Bulk density of the soil	1.5(g/cm^3)
ρ_s	Bulk density of the solid soil particles	2.66 (g/cm^3)
m_v	Volumetric water content (VWC) or moisture	5% or 50%
α'	Soil-type empirically determined constant	0.65
β'	Soil-type empirically determined constant with C clay fraction and S sand fraction	$1.2748 - 0.519S - 0.152C$
β''	Soil-type empirically determined constant	$1.3379 - 0.603S - 0.166C$
ε_s	Relative complex dielectric constant of the mixture of soil and water	$(1.01 + 0.44\rho_s)^2 - 0.062$
τ_w	Relaxation time of water	8×10^{-12}
ε_{w0}	Static dielectric constant of water	80.4 (F m^{-1})
$\varepsilon_{w\infty}$	High-frequency limit of ε'_{fw}	5.0 (F m^{-1})
σ_{eff}	Effective conductivity depending on soil texture	$0.046 + 0.220\rho_b - 0.411S + 0.661C$
ε_0	Permittivity constant of free space	8.854×10^{-12} (F m^{-1})
μ	Magnetic permeability	$4\pi \times 10^{-7}$ (H m^{-1})

expected, we observe that the path losses increase with increasing distance d. Furthermore, the increase in operating frequency f leads to the increase in path loss. This analysis motivates the need to operate at the lowest possible frequencies in the soil medium in all cases, whether with the Peplinski principle or with the combination of Peplinski

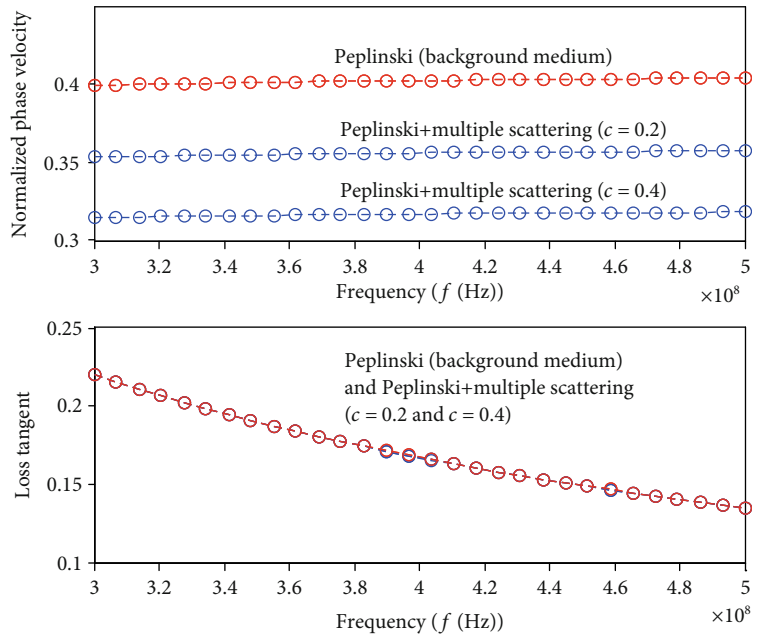


Figure 6: Normalized phase velocity and loss tangent versus frequency at VWC = 5%.

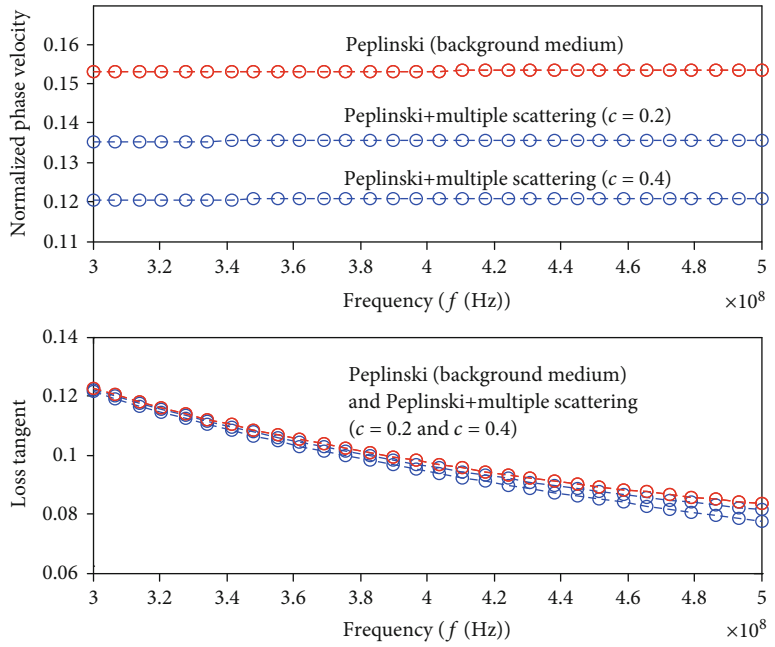


Figure 7: Normalized phase velocity and loss tangent versus frequency at VWC = 50%.

and multiple scattering. Given the above results, a trade-off is required between operating at higher frequencies with a small antenna size but greater path loss and operating at a lower frequency using a bigger antenna but less path loss. An appropriate frequency range between 300 and 900 MHz will be suitable for maintaining small antenna sizes [9]. This will ensure that the sensors remain concealed, a property which is distinctively useful for security applications.

We show the effect of VWC on the path loss for two values, 5% and 50%. The path loss increases with higher proportions of VWC. This effect is particularly important since water content not only depends on the location of the network but also varies during different seasons and should therefore be considered in the design of WUSNs. The comparison between Figures 8(a) and 8(b) leads to the conclusion that it is not necessarily multiple scattering that provokes the high path loss values of $L_p^{Pe+\text{m.s.}}$ as shown in Figure 8(a). For example, in Figure 8(b), at the operating frequency of 868 MHz, $L_p^{Pe+\text{m.s.}} < L_p^{Pe}$. This shows that signal transmission is severely affected in soil containing not only random

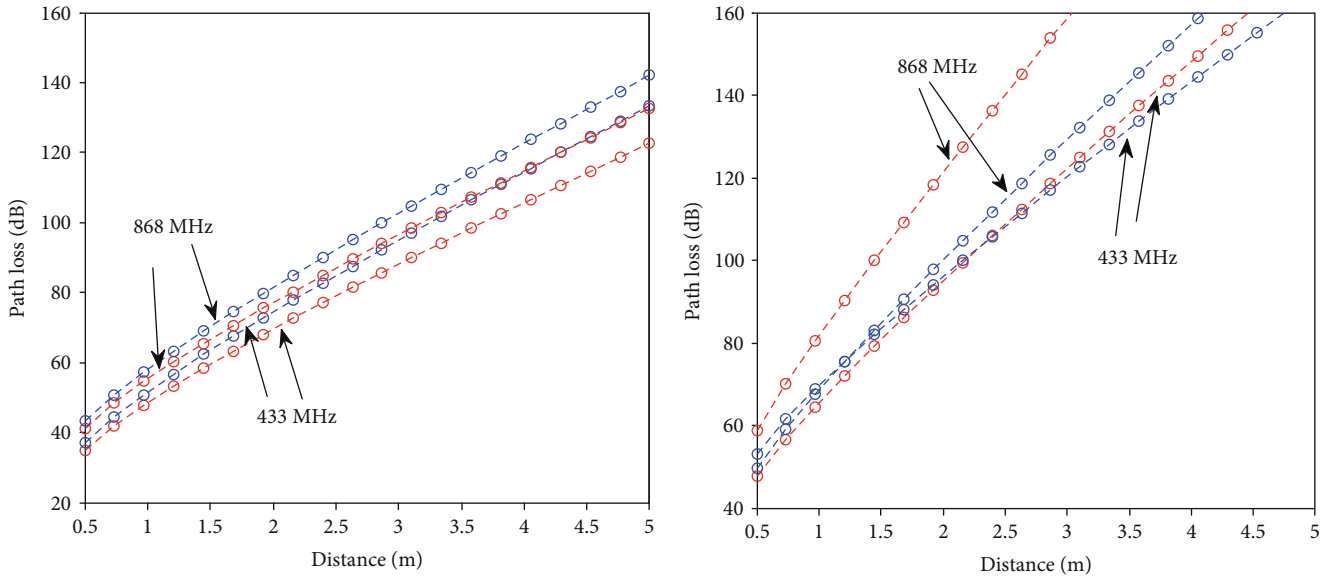


Figure 8: Path loss *vs.* distance and frequency (a) at VWC = 5% and (b) at VWC = 50%.

distribution of spheres (stones) but also another important parameter which is water. We chose a relative permittivity $\varepsilon_p = 3.2$ for the spheres. It must be noted that changing this value modifies the results discussed above. In particular, if $\varepsilon_p = 1$, for example, L_p^{Pe} and $L_p^{Pe+\text{m.s.}}$ become identical as expected.

7. Discussion

In this paper, a new model of the effective wave number that accounts for absorption due to permittivity and multiple scattering occurring in soil because of the presence of buried obstacles such as stones, rocks, or pebbles has been proposed. From the analysis of normalized phase velocity and the loss tangent of wave propagation in soil as shown in Figures 4 and 5, it can be concluded that the wave velocity is less dependent on the concentration and the size of particles in soil. Furthermore, waves propagate faster in the soil medium with the lowest concentration of particles but with a higher loss tangent.

The graphs in Figures 6 and 7 indicate that a rise in the soil moisture causes a decrease of the phase velocity and the decay of the loss tangent, which diverge slightly as the frequency increases. In a nutshell, higher volumetric water content in soil reduces wave velocity leading to very slow wave propagation in soil.

Multiple scattering effects combined with Peplinski's principle enabled us to derive a mixed model of the effective wave number in soil, which accounts for both moisture and particle with a spherical shape. The numerical results show how the integration of the multiple scattering in the analysis modifies the path loss, which is a very important performance indicator for underground communication and adequate for the implementation of wireless underground sensor networks (WUSN).

8. Conclusion

In this paper, we have compared the wave number based upon a multiple scattering model in a dense medium soil with predictions based on the semiempirical model of Peplinski. The soil is assumed to contain a random distribution of scatterers or particles, which makes the estimation of the signal propagation in that medium quite challenging. By combining multiple scattering effect and Peplinski's principle, we derive a mixed model of the effective wave number in soil which accounts for both moisture and stones (of spherical shape). The numerical results show how the integration of the multiple scattering in the analysis modifies the path loss which is an important parameter for any wireless communication system and in the design and conceptualization of wireless underground sensor networks (WUSNs).

Data Availability

The data used to support the simulation results and the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The authors acknowledge the Department of Computer Science at the University of Ghana and the Department of Information Technology, UPSA, for their support. Special thanks are due to Laboratoire Ondes et Milieux complexes (LOMC) UMR CNRS 6294, University of Le Havre, France.

References

[1] D. W. Sambo, A. Forster, and B. O. Yenke, "Wireless underground sensor networks path loss model for precision agriculture (WUSNPLM)," *IEEE Sensors Journal*, vol. 20, no. 10, pp. 5298–5313, 2020.

[2] A. Salam, "Subsurface MIMO: a beamforming design in internet of underground things for digital agriculture applications," *Journal of Sensors and Actuator Networks*, vol. 8, no. 3, p. 41, 2019.

[3] A. Villa-Henriksen, G. T. C. Edwards, L. A. Pesonen, O. Green, and C. A. G. Sørensen, "Internet of things in arable farming: implementation, applications, challenges and potential," *Journal of Biosystems Engineering*, vol. 191, pp. 60–84, 2020.

[4] H. Huang, J. Shi, F. Wang, D. Zhang, and D. Zhang, "Theoretical and experimental studies on the signal propagation in soil for wireless underground sensor networks," *Sensors*, vol. 20, no. 9, article 2580, 2020.

[5] U. Raza and A. Salam, "Wireless underground communications in sewer and stormwater overflow monitoring: radio waves through soil and asphalt medium," *Information*, vol. 11, no. 2, p. 98, 2020.

[6] M. C. Vuran and I. F. Akyildiz, "Channel model and analysis for wireless underground sensor networks in soil medium," *Physical Communication*, vol. 3, no. 4, pp. 245–254, 2010.

[7] K. Arshad, F. Katsriku, and A. Lasebae, "Radiowave VHF propagation modelling in forest using finite elements," in *Proceedings -2006 International Conference on Information and Communication Technologies*, pp. 2146–2149, Damascus, Syria, 2005.

[8] K. Arshad, F. Katsriku, and A. Lasebae, "Modeling obstructions in straight and curved rectangular tunnels by finite element approach," *Journal of Electrical Engineering–Bratislava*, vol. 59, no. 1, pp. 09–13, 2007.

[9] K. Arshad, F. Katsriku, and A. Lasebae, "Effects of different parameters on attenuation rates in circular and arch tunnels," *Piers Online*, vol. 3, no. 5, pp. 607–611, 2006.

[10] H. T. Friis, "A note on a simple transmission formula," *Proceedings of the IRE*, vol. 34, no. 5, pp. 254–256, 1946.

[11] Z. Sun and I. F. Akyildiz, "Magnetic induction communications for wireless underground sensor networks," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 7, pp. 2426–2435, 2010.

[12] X. Tan, Z. Sun, and I. F. Akyildiz, "Wireless underground sensor networks: MI-based communication systems for underground applications," *IEEE Antennas and Propagation Magazine*, vol. 57, no. 4, pp. 74–87, 2015.

[13] A. Salam, M. C. Vuran, and S. Irmak, "Pulses in the sand: impulse response analysis of wireless underground channel," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, San Francisco, CA, USA, 2016.

[14] F. H. Liedmann, C. Holewa, and C. Wietfeld, "The radio field as a sensor-a segmentation based soil moisture sensing approach," in *2018 IEEE Sensors Applications Symposium (SAS)*, Seoul, South Korea, 2018.

[15] N. R. Peplinski, F. T. Ulaby, and M. C. Dobson, "Dielectric properties of soils in the 0.3–1.3-GHz range," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 3, pp. 803–807, 1995.

[16] L. K. Tsang, J. A. Kong, and K.-H. Ding, "Scattering and emission by layered media," in *Scattering of Electromagnetic Waves: Theories and Applications*, pp. 200–229, Wiley, New York, 1st edition, 2000.

[17] L. Tsang, J. A. Kong, K.-H. Ding, and C. O. Ao, "Particle positions for dense media characterizations and simulations," in *Scattering of Electromagnetic Waves: Numerical Simulations*, pp. 403–451, Wiley, New York, 1st edition, 2001.

[18] L. Tsang and J. A. Kong, "Quasi-crystalline approximation in dense medium scatteing," in *Scattering of Electromagnetic Waves: Advanced Topics*, pp. 245–319, Wiley, New York, 1st edition, 2001.

[19] M. C. Vuran, A. Salam, R. Wong, and S. Irmak, "Internet of underground things in precision agriculture: architecture and technology aspects," *Ad Hoc Networks*, vol. 81, no. 1, pp. 160–173, 2018.

[20] I. F. Akyildiz, "Wireless sensor networks in challenged environments such as underwater and underground," in *Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Montreal, QC, Canada, 2014.

[21] M. A. Akkaş and R. Sokullu, "Wireless underground sensor networks: channel modeling and operation analysis in the terahertz band," *International Journal of Antennas and Propagation*, vol. 2015, 12 pages, 2015.

[22] I. F. Akyildiz, Z. Sun, and M. C. Vuran, "Signal propagation techniques for wireless underground communication networks," *Physical Communication*, vol. 2, no. 3, pp. 167–183, 2009.

[23] I. C. Dumitrache, S. I. Caramihai, I. S. Sacala, and M. A. Moisescu, "A cyber physical systems approach for agricultural enterprise and sustainable agriculture," in *Proceedings -2017 21st International Conference on Control Systems and Computer (CSCS)*, Bucharest, Romania, 2017.

[24] M. Richardson Ansah, R. A. Sowah, J. Melià-Seguí, F. A. Katsriku, X. Vilajosana, and W. Owusu Banahene, "Characterising foliage influence on LoRaWAN pathloss in a tropical vegetative environment," *IET Wireless Sensor Systems*, vol. 10, no. 5, pp. 198–207, 2020.

[25] Y. S. Lei, P. Siqueira, and R. Treuhaft, "A dense medium electromagnetic scattering model for the InSAR correlation of snow," *Radio Science*, vol. 110, no. 1, pp. 461–480, 2016.

[26] L. Tsang, J. A. Kong, and R. T. Shin, *Theory of Microwave Remote Sensing*, New York, 1985.

[27] J. K. Percus and G. J. Yevick, "Analysis of classical statistical mechanics by means of collective coordinates," *Physical Review*, vol. 110, no. 1, pp. 1–13, 1958.

[28] L. Li, M. C. Vuran, and I. F. Akyildiz, "Characteristics of underground channel for wireless underground sensor networks," *The 6th Annual Mediterranean Ad Hoc Networking WorkShop*, pp. 92–99, 2007.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8894852, 11 pages
https://doi.org/10.1155/2020/8894852

Research Article

A Random Walk-Based Energy-Aware Compressive Data Collection for Wireless Sensor Networks

Keming Dong,[1] **Chao Chen**,[2] **and Xiaohan Yu**[2]

[1]*School of Information, Yunnan University of Finance and Economics, Kunming 650221, China*
[2]*School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

Correspondence should be addressed to Xiaohan Yu; yuxiaohan188@126.com

Received 28 August 2020; Revised 17 October 2020; Accepted 24 October 2020; Published 30 November 2020

Academic Editor: Xingsi Xue

The energy efficiency for data collection is one of the most important research topics in wireless sensor networks (WSNs). As a popular data collection scheme, the compressive sensing- (CS-) based data collection schemes own many advantages from the perspectives of energy efficiency and load balance. Compared to the dense sensing matrices, applications of the sparse random matrices are able to further improve the performance of CS-based data collection schemes. In this paper, we proposed a compressive data collection scheme based on random walks, which exploits the compressibility of data vectors in the network. Each measurement was collected along a random walk that is modeled as a Markov chain. The Minimum Expected Cost Data Collection (MECDC) scheme was proposed to iteratively find the optimal transition probability of the Markov chain such that the expected cost of a random walk could be minimized. In the MECDC scheme, a nonuniform sparse random matrix, which is equivalent to the optimal transition probability matrix, was adopted to accurately recover the original data vector by using the nonuniform sparse random projection (NSRP) estimator. Simulation results showed that the proposed scheme was able to reduce the energy consumption and balance the network load.

1. Introduction

This paper considers the energy efficiency issue of compressive data collection in wireless sensor networks (WSNs). A WSN is consisted of low-cost, low-power, and energy-constrained sensors which acquires and transmits information to the sink through wireless links [1–5]. In the area of Internet of Things (IoT), a WSN is regarded as a key technology for the data sensing and collection [6, 7]. One of the most important factors that affects the performance of WSNs is the energy limitation of sensors [8, 9]. A sensor will cease to operate if it depletes its battery energy. We intend to design an energy-efficient data collection scheme by applying the compressive sensing (CS) technology and random walks.

A popular approach to the data collection problem is the application of the CS technology [10, 11]. In the CS technology, data are assumed to be sparse or sparse under some basis, which is very appropriate for data in WSNs [12, 13]. The key idea behind CS is that, by exploiting the sparsity of the original data vector, a high dimensional data vector can be reliably recovered from a significantly lower number of measurements. Early works mainly focus on applying the CS technology with a dense sensing matrix [14–16]. Luo et al. [14] proposed the Compressive Data Gathering (CDG) scheme in which the sink collects linear combinations of the original data vector instead of the individual data sample. The sink is able to recover the original data vector through solving an ℓ_1-based convex optimization as long as a sufficient number of linear combinations are collected. Compared with traditional schemes, the CDG scheme not only reduces the energy consumption but also evenly distributes loads across the network. The reference [15] improved the CDG scheme and proposed a hybrid-CS scheme in which data are only encoded at overloaded nodes. This significantly reduces the load of nodes which are far away from the sink. The authors showed that, compared with the CDG scheme, the hybrid-CS scheme can further improve the throughput of networks. Adopting the idea of the CDG scheme, the reference [16] considered not only the energy efficiency but also the delay of data collection. The authors proposed a joint

optimization problem which aims to minimize the delay of data collection with bounded transmissions. The NP-hardness of the joint optimization problem was proved. Thus, the authors proposed an approximation solution which decomposes the joint optimization problem into a forwarding tree construction subproblem and a link-scheduling subproblem.

Without the sacrifice of recovery fidelity, sparse random matrices have been proven to give better energy efficiency than the dense random matrices [17, 18]. Under the CDG framework of WSNs, the sparse random matrix can be either uniform [17–20] or nonuniform [21–24]. In the uniform sparse random matrix, each entry is equal to zero with an identical probability. However, in the nonuniform sparse random matrix, entries in different columns are equal to zero with variational probabilities. Wang et al. [17] proposed a class of uniform sparse random matrices that do not compromise the recovery performance when compared to a Gaussian sensing matrix. In their scheme, each node aggregates one measurement as a linear combination of the original data vector. The sink collects measurements from nodes through different shortest paths. Zheng et al. [18] proposed a random walk-based data collection scheme and provided mathematical foundations from the perspectives of the CS and graph theory. They showed that uniform sparse random matrices which are constructed from the proposed random walk scheme satisfy the expansion property of expander graphs. Singh et al. [19] proposed an On-Demand Explosion-Based Compressive Sensing (ODECS) technology to reduce the required number of measurements for the recovery of data vector by exploiting the rate of change of the data vector. The ODECS technology is able to adapt itself to the occurrence of events. It has very low communication rate when events are absent. Considering problems in existing schemes, such as the semidynamic routing, the nonuniform sampling, and the dependence on global coordinate information, Zhang et al. [20] proposed a dual random walk-based compressive data collection scheme. In the proposed scheme, a dual random walk, which does not rely on coordinate information, was first designed to achieve a uniform sampling. Then, depending on the dual random walk, a dynamic and distributed CDG-based scheme was proposed to enhance the network dynamic adaptability.

Recently, nonuniform sparse random matrices were proved to give similar performance as the uniform sparse random matrices [21–24]. Liu et al. [21] proposed a novel compressive data collection scheme which compresses data under an opportunistic routing. The proposed scheme requires fewer compressed measurements and allows a simpler routing strategy without excessive computation and overheads. Moreover, the authors proposed the nonuniform sparse random projection (NSRP) algorithm to recover the original data vector. They proved that the NSRP-based estimator can achieve the optimal estimation error bound. Considering the large transmission energy consumption and low recovery accuracy problem in traditional schemes, Zhang et al. [22] proposed a ring topology-based compressive sensing data collection scheme. In the proposed scheme, the total number of hops is reduced by a ring topology-based random walk, and the recovery accuracy is improved by the dual compensation-based compressive sensing measurements. Huang and Soong [23] proposed a cost-aware stochastic compressive data collection scheme, where the cost diversity and the stochastic data collection process are considered by using the Markov chain model. The proposed scheme is aimed at minimizing the expected cost of a random walk subjected to constraints on the global degree of randomness and recovery error. Without loss of the recovery accuracy, the proposed scheme not only reduces the expected cost but also prolongs the network lifetime due to the load balance feature. The reference [24] proposed a mobile CDG scheme including a random walk-based algorithm and a kernel-based method for sparsifying sensory data from an irregular deployment. The sensing matrix, which is constructed from the proposed random walk algorithm combined with a kernel-based sparsity basis, was proved to satisfy the restricted isometry property. Moreover, the authors proved that $O(k \log (n/k))$ measurements, which can be collected within $O(k \log (n/k))$ steps, were sufficient for the accurate recovery of k-sparse signals in a network with n nodes.

In this paper, we propose a data collection scheme for WSNs by integrating the compressive sensing technology and random walks. The total amount of energy consumption is reduced by exploiting the compressibility of the original data vector. Measurements are collected along random walks so that the local energy consumption is balanced. Specifically, each measurement is a linear combination of the original data in nodes which occur in a random walk. Each random walk is formulated as follows. Initially, a node except for the sink is selected as the starting node with a probability that is determined by the residual energy of every node in the network. Then, data are forwarded to the sink in a multihop manner. During the transmission process, each node selects the next hop node from its candidate nodes according to a probability distribution that is determined by the residual energy of its candidate nodes. We can model this stochastic process as an absorbing Markov chain. The key problem is that how to determine the transition probabilities of nodes in every random walk. We formulate this problem as an optimization problem which aims to find the optimal transition probability matrix such that the expected cost of a random walk is minimized. The Minimum Expected Cost Data Collection (MECDC) scheme is proposed to iteratively find the optimal transition probability matrix. After obtaining the optimal transition probability matrix, the sink is able to construct an equivalent sensing matrix based on the optimal transition probability matrix. Eventually, by using the NSRP-based estimator [21], the original data vector can be accurately recovered. For the compressive data collection problem, the reference [23] adopted a similar idea as this paper. However, in their scheme, each random walk starts at a fixed node, which results in the rapid energy expenditure of the fixed node. This paper extends the reference [23] mainly in five aspects: (1) the starting node of random walks is variational; (2) nodes' residual energy is considered for the balance of network load; (3) the MECDC scheme along with its distributed realization is proposed; (4) the process of collecting measurements is accelerated by partitioning the

network into layers; (5) computation of optimal transition probability matrix is simplified.

The main contributions of this paper are summarized as follows:

(i) We propose a random walk-based compressive data collection scheme which exploits the compressibility of the original data vector. Random walks are responsible for the collection of measurements. In order to reduce the energy consumption and balance the network load, the residual energy of nodes is considered in the process of data collections

(ii) The absorbing Markov chain model is adopted to characterize the stochastic of a random walk. We formulate an optimization problem to minimize the expected cost of a random walk and propose the MECDC scheme to find the optimal transition probability matrix

(iii) A distributed realization of the MECDC scheme is proposed, where the update of transition probabilities for each node can be obtained only based on the information of its neighbors

(iv) Simulation results are provided to demonstrate that the proposed scheme can both reduce the energy consumption and balance the network load

The remainder of this paper is organized as follows. In Section 2, we present preliminaries of this paper. Next, we introduce the system model and problem formulation in Section 3. The MECDC scheme is proposed in Section 4. In Section 5, we present simulation results. Finally, Section 6 concludes the paper.

2. Preliminaries

We will use boldface letters to denote vectors and matrices. The ith entry of vector $\mathbf{x}$ is denoted by x_i. The entry in the i th row and jth column of matrix $\mathbf{A}$ is denoted by a_{ij}. Denote $[n] \coloneqq \{1, 2, \cdots, n\}$. A vector $\mathbf{x}$ is said to be k-sparse if the number of nonzero entries does not exceed k. Consider the following linear model:

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ $(m \ll n)$ is referred as the *sensing matrix*, each entry y_i in vector $\mathbf{y}$ is referred as a *measurement*, and $\mathbf{x} \in \mathbb{R}^n$ is the data vector to be recovered. It is well known that the Gaussian random matrix can be used as the sensing matrix. When $m \geq O(k \log n)$, a k-sparse data vector can be recovered with high probability via linear programming [25, 26].

It has been shown that sparse random matrices provide similar recovery performance as the Gaussian random matrix [17, 27]. A sparse random matrix is a matrix whose entries are zero with some probability. Importantly, if $a_{ij} = 0$, we will *not* need the data x_j when collecting y_i, because y_i is a linear combination of entries in $\mathbf{x}$. By exploiting the sparsity of the sensing matrix, potential improvement including the reduced energy and data collection delay can be obtained [8, 9]. In this paper, we consider the problem of recovering a compressible data vector by using a nonuniform sparse random matrix. Compressible data vectors can be seen as a subset of sparse data vectors. Specifically, a compressible data vector $\mathbf{x}$ can be represented as $\mathbf{x} = \mathbf{\Psi}\theta$, where Ψ is an $n \times n$ orthonormal basis and θ is a coefficient vector that decays according to the power law [28]. If we rearrange entries of θ according to the magnitude, then the ith largest entry $\theta_{(i)}$ satisfies

$$|\theta_{(i)}| \leq ci^{-1/z}, i \in [n], \tag{2}$$

where c is a constant and z controls the rate of decaying. Throughout this paper, we assume that the data vector $\mathbf{x}$ is compressible in some basis. The best k-term approximation of $\mathbf{x}$ is to keep the largest k coefficients and set the others to zero. Let $\widehat{\theta}_k$ be the coefficient vector of the best k-term approximation of $\mathbf{x}$. Then, we have that [28].

$$\|\mathbf{x} - \hat{\mathbf{x}}_k\|_2 = \|\theta - \widehat{\theta}_k\|_2 \leq \zeta_r ck^{-1/r+1/2}, \tag{3}$$

where $\hat{\mathbf{x}}_k = \mathbf{\Psi}\widehat{\theta}_k$ and ζ_r are constant that only depends on r. For a compressible data vector $\mathbf{x}$, the reference [21] proposed a nonuniform sparse random projection-based estimator which gives comparable recovery performance as the best k -term approximation provided that $m = O(k^2 \log n)$ and entries in $\mathbf{A} \in \mathbb{R}^{m \times n}$ are drawn i.i.d. from the following distribution [17, 21, 23].

$$a_{ij} = \begin{cases} +1, & \text{with probability } \dfrac{\pi_j}{2}, \\ -1, & \text{with probability } \dfrac{\pi_j}{2}, \\ 0, & \text{with probability } 1 - \pi_j, \end{cases} \tag{4}$$

where $0 \leq \pi_j \leq 1$ is a probability. Unlike the uniform sparse random matrices, the probability of being zero for entries in different columns of the nonuniform sparse random matrix varies.

3. System Model and Problem Formulation

3.1. Network Model. In this paper, we consider a multihop wireless sensor network consisting of n nodes with node n being the sink. Sensors are randomly deployed in a sensing field to sense the surrounding environment and then periodically report readings to the sink through multihop transmissions. Define x_i^t as the reading of sensor i at time instant t. The sink aims to collect data $\mathbf{x}^t = [x_1^t, x_2^t, \cdots, x_{n-1}^t]$ for different time instants. Previous works [17, 28] have shown that most natural classes of signals, such as smooth signals with bounded derivatives and bounded variation signals, are compressible in some transform domain. As stated in the previous section, we assume that the data $\mathbf{x}^t$ is compressible.

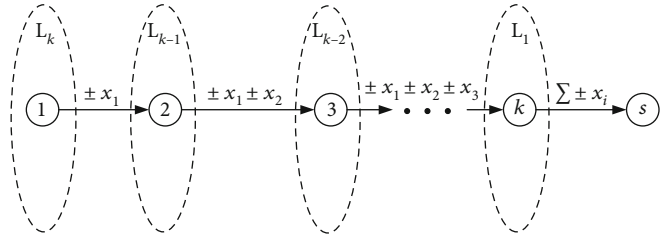


Figure 1: The example of collecting a measurement through a random walk. The random walk starts at node $1 \in L_k$ and ends at the sink.

Without loss of generality, we assume that sensors are randomly deployed in a unit square, and each sensor is equipped with an identical battery with the initial power E_0. Any two nodes are able to communicate with each other if the Euclidean distance between these two nodes is no more than the communication range R. The WSN is modeled as a connected graph $G = (V, E)$ with $V = [n]$ the set of nodes including the root/sink n and E the set of edges/wireless links. Each edge is associated with a weight which is related to the residual energy of nodes. Specifically, we define w_{ij}, the ijth entry of the weight matrix $\mathbf{W} \in \mathbb{R}^{(n-1)\times(n-1)}$, as the weight of edge (i, j) representing the cost of transmitting data from node i to node j. Note that we omit edges related to the sink. Suppose each node knows information of its neighbors. Expect for the sink, we partition nodes into layers $L_k, k = 1, \cdots, T$, where L_k is consisted of the nodes at distance k from the sink. For any node $i \in L_k$, its neighbors are divided into two disjoint sets: the successors set $S_i := \{j \mid (i, j) \in E, j \in L_{k-1}\}$ and the predecessors set $D_i := \{j \mid (i, j) \in E, j \in L_{k+1}\}$. Let $E_r(i)$ be the residual energy of node i. At the beginning of data collection, each node contains an identical initial energy E_0.

3.2. Opportunistic Routing. In this subsection, we describe how measurements are collected by the sink through random walks. The process of collecting measurements can be modeled as a discrete absorbing Markov chain [29] with the state set $\{s_1, s_2, \cdots, s_n\}$ and the transition probability matrix $\mathbf{P}$. Each node in the network corresponds to a state in the discrete absorbing Markov chain. Specifically, we assume that node $i \in [n]$ corresponds to the state s_i, and s_n is the absorbing state. The ijth entry in $\mathbf{P}$, i.e., the state transition probability p_{ij}, corresponds to the probability that the data is transmitted from node i to node j.

Our goal is to collect m measurements through m random walks. Each measurement corresponds to a random walk that starts from a randomly selected node and ends at the sink. Figure 1 shows the process of collecting a measurement, say the jth measurement y_j, which corresponds to the jth random walk. Initially, node 1 in layer L_k is chosen as the starting node. Then, it transmits data $+x_1$ or $-x_1$ to the randomly selected node $2 \in S_1$. Note that $S_1 \subseteq L_{k-1}$. Subsequently, node 2 adds or subtracts the received value to its own data and transmits the result, i.e., $\pm x_1 \pm x_2$, to a randomly selected node, say node $3 \in S_2$. The above process is repeated until the sink receives the measurement $y_j = \sum_{i=1}^{k} \pm x_i$. We can observe that the length of jth random walk is exactly the layer index of the starting node.

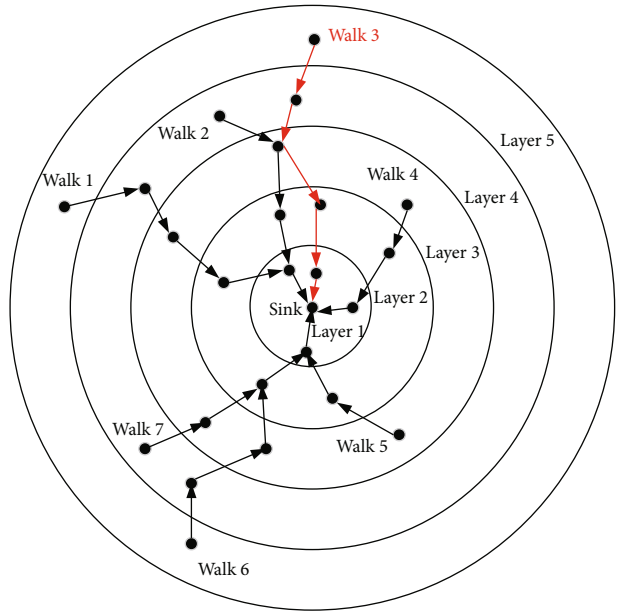


Figure 2: The process of collecting seven measurements. Measurements are collected through random walks. Each random walk starts from a randomly selected node.

In general, the process of collecting each measurement starts at a randomly chosen node. In this paper, a node $i (i \neq n)$ is selected as the starting node with probability p_i. Then, node i randomly selects a successor according to a certain probability distribution and subsequently transmits its compressed data to the selected successor. After receiving data, the selected successor adds or subtracts its own data to the received data and transmits the result towards the sink. The process is repeated until the sink collects every measurement. Figure 2 shows the process of collecting seven measurements. In this figure, nodes are partitioned into layers based on its length to the sink, and there are five layers in Figure 2.

3.3. The Transition Probability Matrix and the Sensing Matrix. The long-term behavior of random walks is closely related to the sensing matrix under the CS framework. In order to see this, let us write the transition probability matrix in the canonical form [29].

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & * \\ \mathbf{0} & 1 \end{pmatrix}, \tag{5}$$

where $\mathbf{Q} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the transition probability matrix of transient states and $\mathbf{0}$ is a row vector with zero entries. It is well known that the fundamental matrix, i.e., $\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}$, represents the long-term behavior of the discrete absorbing Markov chain. Specifically, the ijth entry of $\mathbf{F}$, f_{ij}, gives the expected number of times that the chain is in the transient state s_j, if it is started in the transient state s_i. In other words, f_{ij} is the expected number of occurrence of node j if the random walk starts at node i. In our formulations, every node occurs at most once in a random walk. Therefore, f_{ij} represents the probability of a random walk that passes the node j if it is started at the node i. Furthermore, excepting for the sink n, node i is selected as the starting node with probability p_i. Then, we have that

$$\pi_j = \sum_{i=1}^{n-1} p_i f_{ij}, \tag{6}$$

where π_j, the probability of node j in a random walk, is referred as the compression probability. As stated in the references [21, 23], π_j is exactly the nonzero probability of entries in the jth column of the sensing matrix $\mathbf{A}$.

3.4. Problem Formulation. The energy efficiency is the key issue in this paper. We intend to decrease the energy consumption of data collection and meanwhile balance the load of sensors. Since the opportunistic routing is a stochastic method, a natural idea to minimize the expected cost of a random walk. Specifically, we define c_i, the ith entry in the vector $\mathbf{c} \in \mathbb{R}^{n-1}$, as the expected cost of a random walk if node i is selected as the starting node. The goal is to minimize the expected cost of a random walk which is given by

$$\sum_{i=1}^{n-1} p_i c_i. \tag{7}$$

Furthermore, for any c_i, we have that

$$c_i = \sum_{j \in S_i} q_{ij}\left(w_{ij} + c_j\right). \tag{8}$$

An immediate observation is that $c_i > c_j$ if $i \in L_k, j \in L_r$ with $k > r$. In other words, the expected cost of a random walk with the starting point in a high layer is more than that in a low layer.

Similar to the reference [23], we introduce the concept of randomness for data collection in order to avoid the vulnerability to attack and load unbalance. Specifically, by using the Shannon entropy [30], the local randomness of node i is denoted by

$$h_i = -\sum_{j \in S_i} q_{ij} \log q_{ij}. \tag{9}$$

Obviously, the uniform distribution achieves the maximum local randomness for each node. Let us consider a random walk with starting node k. The randomness of such a random walk is defined as the sum of weighted local randomness of nodes in the random walk:

$$H_k = \sum_{i=1}^{n-1} f_{ki} h_i = -\sum_{i=1}^{n-1} f_{ki} \sum_{j \in S_i} q_{ij} \log q_{ij}. \tag{10}$$

Eventually, the expected randomness of a random walk is given by

$$H = \sum_{k=1}^{n-1} p_k H_k = -\sum_{k=1}^{n-1} p_k \sum_{i=1}^{n-1} f_{ki} \sum_{j \in S_i} q_{ij} \log q_{ij}. \tag{11}$$

The expected randomness of a random walk measures the uncertainty of the measurement that is collected through this random walk.

Recall that the goal is to estimate the transition probability matrix $\mathbf{Q}$ such that the expected cost of a random walk is minimized. Specifically, given the expected randomness of any random walk H and the probabilities of each node being the starting node $\mathbf{p} = [p_1, p_2, \cdots, p_{n-1}]^T$, the problem can be formulated as follows.

$$\min_{\mathbf{Q}} \sum_{i=1}^{n-1} p_i c_i, \tag{12}$$

$$s.t. \sum_{j \in S_i} q_{ij}\left(w_{ij} + c_j\right) = c_i, i \in [n-1], \tag{13}$$

$$-\sum_{k=1}^{n-1} p_k \sum_{i=1}^{n-1} f_{ki} \sum_{j \in S_i} q_{ij} \log q_{ij} = H, \tag{14}$$

$$0 \le q_{ij} \le 1, \tag{15}$$

$$\sum_{j \in S_i} q_{ij} = 1, i \in [n-1], \tag{16}$$

where constraint (13) shows how to compute the cost of a random walk with a given starting node, constraint (14) guarantees the uncertainty of the collected measurements, and constraint (16) states that the sum of the probabilities of selecting successors must be one.

In order to save the energy consumption and balance network loads, we relate the energy efficiency issue to the weight of edges and p_i's. The idea is to assign smaller edge weight and larger starting probability to nodes that contain more residual energy. Specifically, let w_{ij} and p_i be functions of r_i where $r_i := E_r(i)/E_0$ is defined as the proportion of the residual energy of node i normalized by the initial energy E_0. Suppose the starting probability of node i in a random walk is proportional to r_i, i.e., $p_i = \alpha r_i$, where α is a constant. In order to calculate the constant α, let us recall that the sink needs to collect m measurements so that the data vector can be precisely recovered. This means that m random walks are required for the data recovery. Since p_i is also the expected number of random walks that starts at node i, we have that $\sum_{i=1}^{n-1} p_i = m$. Therefore, the constant α is given by

$$\alpha = \frac{1}{m}\sum_{i=1}^{n-1} r_i. \tag{17}$$

The similar idea is also applied to the computation of edge weights. For a node i, we assign larger weight to the edge which is connected to the successor with smaller proportion of the residual energy. The weight of transmitting data from node i to node j is defined as

$$w_{ij} = \frac{1}{r_j}\sum_{j\in S_i} r_j, j \in S_i. \tag{18}$$

4. Minimum Expected Cost Data Collection

In this section, we propose a network layer-based Minimum Expected Cost Data Collection (MECDC) scheme by using the absorbing Markov chain model. The MECDC scheme is consisted of two phases. In the Phase I, the transition probability matrix **Q** is calculated, and the sensing matrix **A** is constructed based on **Q**. In the Phase II, measurements are collected by applying random walks with the transition probability matrix **Q**. After receiving enough number of measurements, the sink is able to recover the original data vector by using the NSRP decoder with the sensing matrix **A** [21, 23].

4.1. Solution of the Optimization Problem. Let us first discuss how to derive the transition probability matrix **Q**. By leveraging the idea in the reference [23], we apply the Lagrange multiplier method to iteratively update transition probabilities. The Lagrange for the optimization problem is given by

$$L = \sum_{i=1}^{n-1} p_i c_i + \sum_{i=1}^{n-1} \lambda_i \left[c_i - \sum_{j\in S_i} q_{ij}(w_{ij} + c_j)\right] + \sum_{i=1}^{n-1} \mu_i \left(\sum_{j\in S_i} q_{ij} - 1\right) + \eta\left[\sum_{k=1}^{n-1} p_k \sum_{i=1}^{n-1} f_{ki} \sum_{j\in S_i} q_{ij} \log q_{ij} + H\right], \tag{19}$$

where λ_i, μ_i, and η are the Lagrangian multipliers.

By setting $\partial L/\partial q_{kl} = 0$, we have that

$$-\lambda_k(w_{kl} + c_l) + \eta(\log q_{kl} + 1)\sum_{i=1}^{n-1} p_i f_{ik} + \mu_k + \eta\sum_{i=1}^{n-1} p_i \sum_{j=1}^{n-1} \frac{\partial f_{ij}}{\partial q_{kl}} h_j = 0, \tag{20}$$

where $h_j = -\sum_{k\in S_j} q_{jk} \log q_{jk}$. After some simple manipulations, we obtain that

$$q_{kl} = \exp\left\{\frac{\lambda_k(w_{kl} + c_l)}{\eta\beta_k} - \frac{\xi_{kl}}{\beta_k} - \frac{\mu_k}{\eta\beta_k} - 1\right\}, \tag{21}$$

where $\beta_k = \sum_{i=1}^{n-1} p_i f_{ik}$ and

$$\xi_{kl} = \sum_{i=1}^{n-1} p_i \sum_{j=1}^{n-1} \frac{\partial f_{ij}}{\partial q_{kl}} h_j. \tag{22}$$

Applying equation (21) for $l \in S_k$ to the fact that $\sum_{l\in S_k} q_{kl} = 1$, we have that

$$\exp\left\{-\frac{\mu_k}{\eta\beta_k} - 1\right\} = \left(\sum_{l\in S_k} \exp\left\{\frac{\lambda_k(w_{kl} + c_l)}{\eta\beta_k} - \frac{\xi_{kl}}{\beta_k}\right\}\right)^{-1}. \tag{23}$$

Substituting equation (23) into equation (21), we obtain that

$$q_{kl} = \frac{\exp\{(\lambda_k(w_{kl} + c_l)/\eta\beta_k) - \xi_{kl}/\beta_k\}}{\sum_{l\in S_k} \exp\{(\lambda_k(w_{kl} + c_l)/\eta\beta_k) - \xi_{kl}/\beta_k\}}. \tag{24}$$

In order to update q_{kl} for a given **Q**, we need to compute parameters λ_k and ξ_{kl}. Setting $\partial L/\partial c_k = 0$, we have that

$$\lambda_k = \begin{cases} -p_k, & if k \in L_t, \\ -p_k - \sum_{r\in D_k} \lambda_r q_{rk}, & \text{otherwise.} \end{cases} \tag{25}$$

Thus, the Lagrange multiplier λ_k can be computed layers by layers.

Next, we compute $\xi_{kl} = \sum_{i=1}^{n-1} p_i \sum_{j=1}^{n-1} (\partial f_{ij}/\partial q_{kl}) h_j$. Denote $L(i)$ the layer of node i. For different nodes i, j, and k, three cases may occur: (1) $L(k) > L(i) > L(j)$; (2) $L(i) > L(j) > L(k)$; (3) $L(i) > L(k) > L(j)$. Note that $f_{ij} = 0$ if $L(i) \le L(j)$. We observe that $\partial f_{ij}/\partial q_{kl} = 0$ if cases (1) and (2) occur. Under the case (3), we have that

$$\frac{\partial f_{ij}}{\partial q_{kl}} = f_{ik} f_{lj}. \tag{26}$$

By substituting equation (26) into equation (22), we obtain that

$$\xi_{kl} = \sum_{i=1}^{n-1} p_i f_{ik} \sum_{j=1}^{n-1} f_{lj} h_j. \tag{27}$$

Given a guess of **Q**, transition probabilities can be updated based on equation (24). Note that it is impossible to obtain an analytical expression of the Lagrange multiplier η [23]. It controls the degree of randomness of a random walk. Larger value of η implies larger degree of randomness. Algorithm 1 shows how to compute the transition probability matrix **Q** iteratively. In line 5 of the Algorithm 1, ε represents the threshold of the stopping criterion.

The sensing matrix **A** can be constructed based on **Q**. Given **Q**, the fundamental matrix is given by $\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}$. Then, each entry in **A** is identically and independently drawn from the following distribution

1 **Input** the graph $G = (V, E)$, the weight matrix $\mathbf{W}$, the starting probabilities of nodes $p_1, p_2, \cdots, p_{n-1}$ and the randomness of random walks η.
2. Compute layers $L_1, L_2, \cdots, L_T$, the successors set S_i and the predecessors set D_i for every node $i \in [n-1]$.
3. **Output** an estimator of $\mathbf{Q}$.
4. **Initialize** the step index $t = 0$ and $\mathbf{Q} = \mathbf{Q}^0$ such that.
$q_{ij}^0 = \{(1/|S_i|), if\, i \notin L_1, j \in S_i, 0, otherwise,$
where q_{ij}^0 is the ij-th entry of $\mathbf{Q}^0$.
5. **while** $\max_{q,l} |q_{kl}^t - q_{kl}^{t-1}/q_{kl}^{t-1}| \geq \varepsilon$ **do**
6. Compute $\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-1}$.
7. Compute $c_i = \sum_{j \in S_i} q_{ij}(w_{ij} + c_j)$ for any $i \in [n-1]$.
8. Compute λ_k based on equation (25) in a layer-by-layer manner.
9. Compute $\beta_j = \sum_{i=1}^{n-1} p_i f_{ij}$ for any $j \in [n-1]$.
10. Compute $h_j = -\sum_{k \in S_j} q_{jk} \log q_{jk}$ for any $j \notin L_1$.
11. Compute ξ_{kl} based on equation (27) for any $k \in [n-1], l \in S_k$.
12. Update q_{kl}^t based on equation (24) for any $k \in [n-1], l \in S_k$.
13. Update the step index $t = t + 1$.
14. end while

Algorithm 1: Iteratively solve for transition probabilities.

$$a_{ij} = \begin{cases} +1, & \text{with probability } \dfrac{\pi_j}{2}, \\ -1, & \text{with probability } \dfrac{\pi_j}{2}, \\ 0, & \text{with probability } 1 - \pi_j, \end{cases} \tag{28}$$

where $\pi_j = \sum_{i=1}^{n-1} p_i f_{ij}$.

After obtaining the transition probability matrix $\mathbf{Q}$, Phase II collects measurements through random walks. Except for the sink, any node i starts a random walk with probability p_i. Then, packets are transmitted towards the sink in a layer-by-layer manner as stated in Section 3.2. Given $\mathbf{A}$, $m = O(k^2 \log n)$ measurements are sufficient for the sink to recover the original data vector by using the NSRP-based estimator [21, 23] .

4.2. Distributed Realization of MECDC. In this subsection, we show that the update of transition probabilities can be realized locally and distributively. In other words, q_{kl} can be computed only using information of neighbors. In order to see this, let us consider the node k. Suppose each node knows the probability of being the starting node. Then, all of the parameters which are required to update q_{kl} can be computed based on the neighboring nodes as follows.

(i) λ_k. From the equation (25), λ_k can be calculated by using the information of predecessors. Specifically, $\lambda_k = -p_k$ for any $k \in L_T$. Then, λ_k for any $k \in L_i$ can be computed based on nodes in $D_k \subseteq L_{i+1}$. In such a manner, the values of λ_k can be computed layers by layers

(ii) β_k. Similar to λ_k, the value of β_k can be computed based on predecessors of node k. Recall that $\beta_k = \sum_{i=1}^{n-1} p_i f_{ik}$, which can be rewritten as

$$\beta_k = \begin{cases} 0, & if\, k \in L_T, \\ p_k + \displaystyle\sum_{i \in D_k} \beta_i q_{ik}, & \text{otherwise.} \end{cases} \tag{29}$$

Therefore, starting from the layer L_T, the values of β_k can be obtained layers by layers.

(i) h_k. Recall that $h_k = -\sum_{l \in S_k} q_{kl} \log q_{kl}$, which can be computed based on successors of node k

(ii) ξ_{kl}. Denote $g_l = \sum_{j=1}^{n-1} f_{lj} h_j$. In order to obtain ξ_{kl}, we first compute g_l for each $l \in [n-1]$. Based on the information of successors, we obtain that $g_l = \sum_{j \in S_l} q_{lj} g_j$. Thus, starting from the layer L_1, the parameter g_l can be computed layers by layers as follows:

$$g_l = \begin{cases} 0, & if\, l \in L_1, \\ h_l, & if\, l \in L_2, \\ \displaystyle\sum_{j \in S_l} q_{lj} g_j, & \text{otherwise.} \end{cases} \tag{30}$$

Based on the values of g_l and the equation (27), we have that $\xi_{kl} = \beta_k g_l$.

In summary, the computation of transition probabilities can be realized distributively by saving the information of parameters λ_k, β_k, h_k, and g_k in every node.

5. Simulation Results

In this section, we numerically evaluate the performance of the proposed scheme with the baseline scheme. Suppose n nodes are uniformly and randomly deployed in a unit square area. The sink is located at the top right corner. There exists

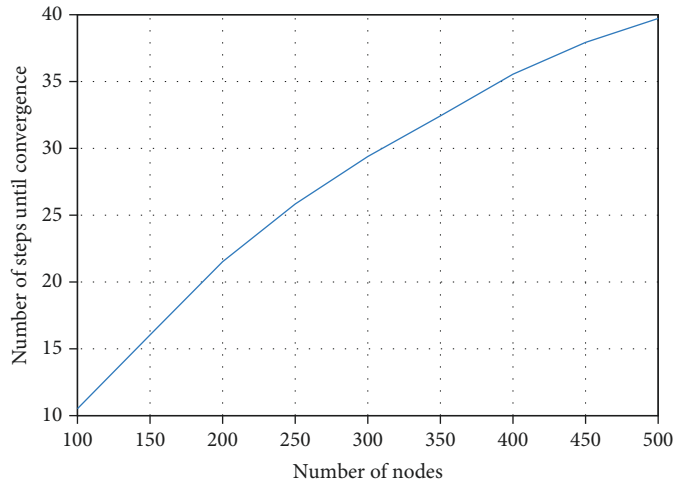


Figure 3: Number of iterations until the Algorithm 1 converges versus the number of nodes in the network. The threshold of the stopping criterion is set to $\varepsilon = 0.1$.

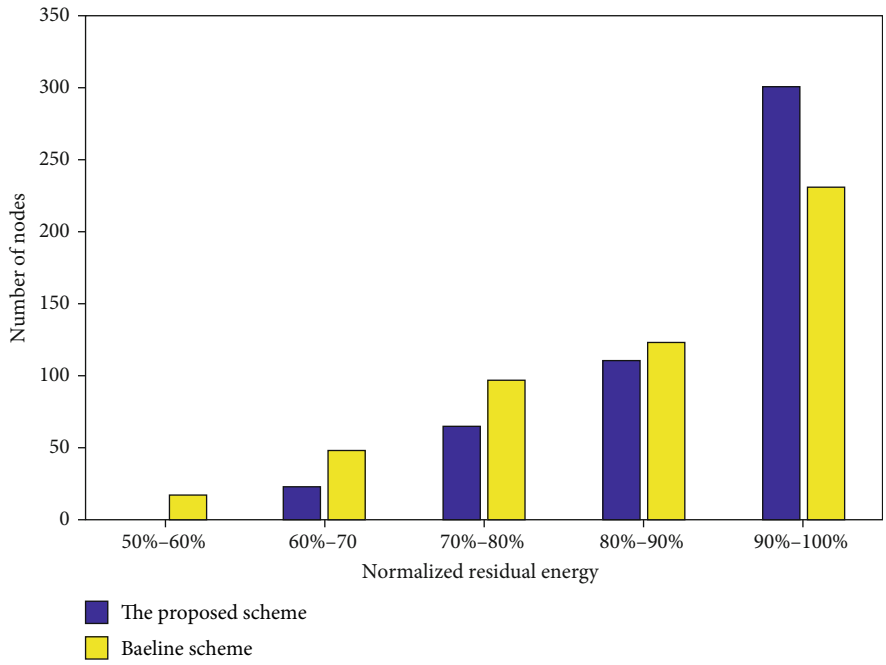


Figure 4: Distribution of the normalized residual energy of nodes in the network.

an edge between two nodes if the distance between these two nodes is not greater than the communication range 0.2. We assume that $m = k^2 \log n$ measurements are required to recover the data vector. The sparsity of the data vector is set to $k = 5$. Initially, each node is equipped with an identical battery that contains 100 joules of the energy. For simplicity, we assume that a packet transmission consumes 0.1 joules of the energy. In the baseline scheme, except for the sink, each node is selected as the starting node of a random walk with probability $p = m/(n-1)$. In a random walk, each node transmits data to its successors with an identical probability, i.e., $q_{ij} = 1/|S_i|$ for any $j \in S_i$. In the proposed scheme, we first compute the starting probability of every node and the transition probability matrix $\mathbf{Q}$ at the beginning of collecting every sample. Then, measurements are collected through random walks with the obtained parameters.

Let us first look at the convergence speed of Algorithm 1. Figure 3 shows the number of iterations until the proposed algorithm converges when the network size increases. In Figure 3, we set the threshold of the stopping criterion $\varepsilon = 0.1$. We observe that Algorithm 1 converges very fast when the network size is not large. The convergence rate increases as the network size increases. One possible reason is that the candidate edge (k, l) with $|(q_{kl}^t - q_{kl}^{t-1})/q_{kl}^{t-1}|$ achieving the maximum value increase as the network size increases. Furthermore, we observe that the slope of the convergence rate

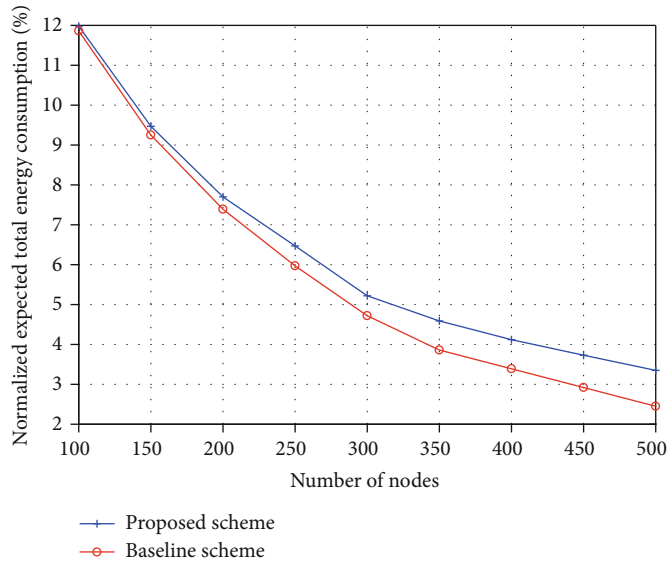


Figure 5: Comparison of the expected total energy consumption between the proposed scheme and the baseline scheme.

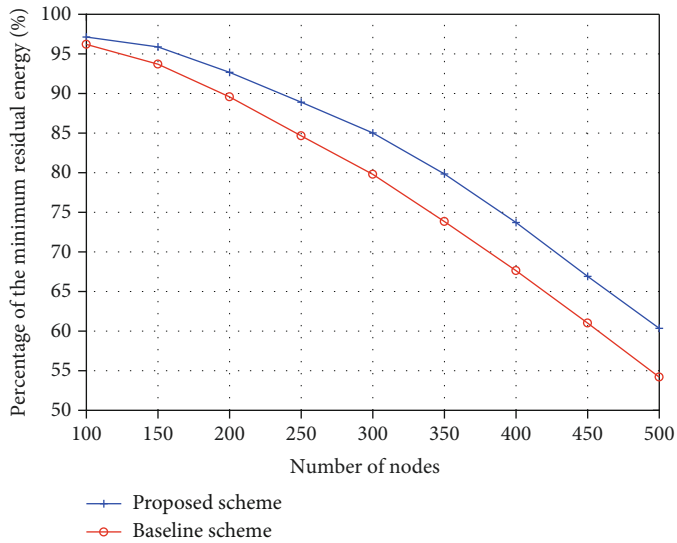


Figure 6: Comparison of the minimum residual energy among nodes when the number of nodes increases.

curve decreases as the network size increases. This implies that there may exist an upper bound for the convergence rate of Algorithm 1.

Next, we compare the energy efficiency between the proposed scheme and the baseline scheme. Figure 4 shows the distributions of the normalized residual energy of nodes in the network. In Figure 4, we set the network size to $n = 500$, and the residual energy is obtained after 50 samples are collected. Note that the residual energy is normalized based on the initial energy. We observe that the residual energy of nodes in both of the two schemes mainly concentrates on the interval 90%-100%. This is because CDG-based schemes can balance network loads. However, the number of nodes with large residual energy in the proposed scheme is more than that in the baseline scheme. This demonstrates that the proposed scheme is able to further reduce the energy consumption and balance the network loads.

Figure 5 compares the normalized expectation of the total energy consumption between the proposed scheme and the baseline scheme. In Figure 5, the residual energy is computed after 50 samples are collected. The expectation of the total energy consumption is normalized based on the initial total energy of nodes in the whole network. We first observe that, for a fixed number of nodes, the normalized expected total energy consumption in the proposed scheme is smaller than that in the baseline scheme. This demonstrates that the proposed scheme is able to reduce the total energy consumption by considering the residual energy of nodes and optimizing the transition probability matrix. Another observation is that, as the number of nodes increases, the normalized

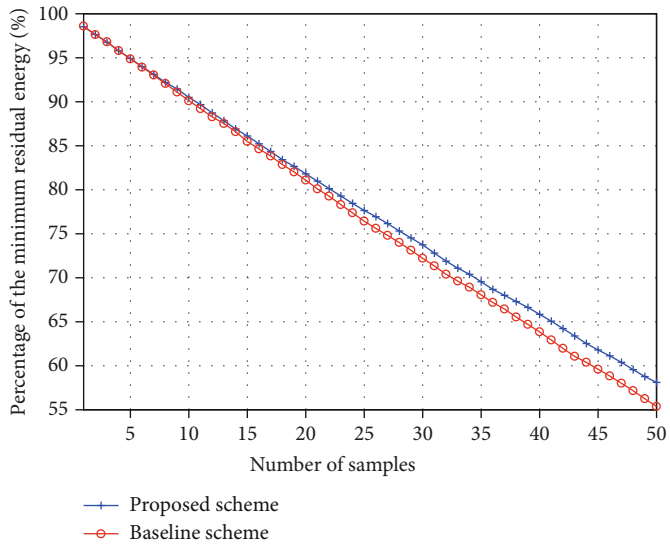


Figure 7: Comparison of the minimum residual energy among nodes when the number of collected samples increases.

expected total energy consumption decreases in both of the two schemes. This implies that collecting a fixed number of samples consumes less normalized total energy for large-scale networks. In other words, the proposed scheme is more suitable for large-scale networks. Finally, we observe that the gap of the normalized expected total energy consumption between the proposed scheme and the baseline scheme increases as the number of nodes increases. This also implies that the proposed scheme performs better in large-scale networks.

Next, let us consider the minimum residual energy of nodes in the network. Large minimum residual energy implies balanced load of the network. Figure 6 compares the minimum residual energy of nodes between the proposed scheme and the baseline scheme. The residual energy is computed after 50 samples are collected. Similar to Figure 5, the residual energy is normalized based on the initial energy. A direct observation is that the minimum residual energy in the proposed scheme is larger than that in the baseline scheme. This demonstrates that the proposed scheme is able to balance the network load. Another observation is that the gap between the proposed scheme and the baseline scheme increases as the number of nodes increases, which suggests that the proposed scheme is more suitable for large-scale networks.

Figure 7 compares the minimum residual energy of nodes when the number of collected samples increases. In Figure 7, we set the number of nodes $n = 500$. In order to collect a sample/data vector, the sink needs to collect m measurements so that the data vector can be precisely recovered. The residual energy is normalized based on the initial energy. We observe that, for a fixed number of samples, the minimum residual energy of the proposed scheme is larger than that of the baseline scheme. This is because the proposed scheme is able to balance loads of the network. Furthermore, the gap of the minimum residual energy between the proposed scheme and the baseline scheme increases as the number of collected samples increases. This demonstrates that the proposed scheme is more suitable for long-running networks.

6. Conclusions

In this paper, we studied the data collection problem in WSNs. Random walks and the compressive sensing technology with nonuniform sparse random matrices are adopted to collect measurements. Each measurement is collected through a random walk which is modeled as an absorbing Markov chain. By exploiting the residual energy of nodes, we formulate the process of collecting measurements as an optimization problem, which seeks to find optimal transition probabilities of nodes so that the expected cost is minimized. An iterative method, which is referred as the Minimum Expected Cost Data Collection (MECDC) scheme, is proposed to solve this optimization problem and collect measurements. A distributed realization of MECDC, where only local information is needed in the collection of measurements, is proposed. Simulation results show that the proposed scheme not only reduces the energy consumption but also balances the network loads.

Abbreviations

WSNs: Wireless sensor networks
CS: Compressive sensing
MECDC: Minimum Expected Cost Data Collection
IoT: Internet of Things
CDG: Compressive data gathering
NSRP: Nonuniform sparse random projection
ODECS: On-Demand Explosion-Based Compressive Sensing.

Data Availability

Data settings can be found in the draft.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61701441 and 61801427 and 2020 Undergraduate Training Program for Innovation and Entrepreneurship of Yunnan Province (202010689042).

References

[1] X. Xue and J. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, p. 2056, 2020.

[2] B. Rashid and M. H. Rehmani, "Applications of wireless sensor networks for urban areas: a survey," *Journal of Network and Computer Applications*, vol. 60, pp. 192–219, 2016.

[3] 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.

[4] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[5] 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.

[6] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in internet of things: a survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 1–27, 2018.

[7] S. Vashi, J. Ram, J. Modi, S. Verma, and C. Prakash, "Internet of things (IoT): a vision, architectural elements, and security issues," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 1–27, Palladam, India, 2017.

[8] X. Yu and S. J. Baek, "Joint routing and scheduling for data collection with compressive sensing to achieve order-optimal latency," *International Journal of Distributed Sensor Networks*, vol. 13, no. 10, 2017.

[9] X. Yu and S. J. Baek, "Energy-efficient collection of sparse data in wireless sensor networks using sparse random matrices," *ACM Transactions on Sensor Networks*, vol. 13, no. 3, pp. 1–36, 2017.

[10] P. Zhang, S. Wang, K. Guo, and J. Wang, "A secure data collection scheme based on compressive sensing in wireless sensor networks," *Ad Hoc Networks*, vol. 70, pp. 73–84, 2018.

[11] M. Tuan Nguyen, K. A. Teague, and N. Rahnavard, "CCS: energy-efficient data collection in clustered wireless sensor networks utilizing block-wise compressive sensing," *Computer Networks*, vol. 106, pp. 171–185, 2016.

[12] S. Li, L. Da Xu, and X. Wang, "Compressed sensing signal and data acquisition in wireless sensor networks and Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2177–2186, 2013.

[13] M. A. Razzaque and S. Dobson, "Energy-efficient sensing in wireless sensor networks using compressed sensing," *Sensors*, vol. 14, no. 2, pp. 2822–2859, 2014.

[14] C. Luo, F. Wu, J. Sun, and C. W. Chen, "Compressive data gathering for large-scale wireless sensor networks," in *MobiCom '09: Proceedings of the 15th annual international conference on Mobile computing and networking*, pp. 145–156, Beijing China, 2009.

[15] J. Luo, L. Xiang, and C. Rosenberg, "Does compressed sensing improve the throughput of wireless sensor networks?," in *2010 IEEE International Conference on Communications*, pp. 1–6, Cape Town, South Africa, 2010.

[16] D. Ebrahimi and C. Assi, "On the interaction between scheduling and compressive data gathering in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2845–2858, 2016.

[17] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," in *2007 6th International Symposium on Information Processing in Sensor Networks*, pp. 331–339, Cambridge, MA, USA, 2007.

[18] H. Zheng, F. Yang, X. Tian, X. Gan, X. Wang, and S. Xiao, "Data gathering with compressive sensing in wireless sensor networks: a random walk based approach," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 1, pp. 35–44, 2015.

[19] V. K. Singh, S. Verma, and M. Kumar, "ODECS: an on-demand explosion-based compressed sensing using random walks in wireless sensor networks," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2466–2475, 2019.

[20] P. Zhang and J. Wang, "On enhancing network dynamic adaptability for compressive sensing in WSNs," *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8450–8459, 2019.

[21] X. Y. Liu, Y. Zhu, L. Kong et al., "*CDC*: compressive data collection for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 8, pp. 2188–2197, 2015.

[22] P. Zhang, J. Wang, and K. Guo, "Compressive sensing and random walk based data collection in wireless sensor networks," *Computer Communications*, vol. 129, pp. 43–53, 2018.

[23] J. Huang and B. H. Soong, "Cost-aware stochastic compressive data gathering for wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1525–1533, 2019.

[24] H. Zheng, W. Guo, and N. Xiong, "A kernel-based compressive sensing approach for mobile data gathering in wireless sensor network systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2315–2327, 2018.

[25] M. Rani, S. B. Dhok, and R. B. Deshmukh, "A systematic review of compressive sensing: concepts, implementations and applications," *IEEE Access*, vol. 6, pp. 4875–4894, 2018.

[26] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: a compressive sensing approach," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 40–58, 2018.

[27] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 937–947, 2010.

[28] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best $_k$-term approximation," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 211–231, 2009.

[29] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, American Mathematical Society, Providence, RI, USA, 2012.

[30] M. Saerens, Y. Achbany, F. Fouss, and L. Yen, "Randomized Shortest-Path problems: two related models," *Neural Computation*, vol. 21, no. 8, pp. 2363–2404, 2009.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8854649, 13 pages
https://doi.org/10.1155/2020/8854649

Research Article

Air Pollution Concentration Forecast Method Based on the Deep Ensemble Neural Network

Canyang Guo, Genggeng Liu, and Chi-Hua Chen

College of Mathematics and Computer Sciences, Fuzhou University, No. 2 Xue Yuan Road, University Town, Fuzhou, Fujian 350116, China

Correspondence should be addressed to Genggeng Liu; liugenggeng@fzu.edu.cn and Chi-Hua Chen; chihua0826@gmail.com

Received 28 August 2020; Revised 18 September 2020; Accepted 21 September 2020; Published 5 October 2020

Academic Editor: Xingsi Xue

The global environment has become more polluted due to the rapid development of industrial technology. However, the existing machine learning prediction methods of air quality fail to analyze the reasons for the change of air pollution concentration because most of the prediction methods take more focus on the model selection. Since the framework of recent deep learning is very flexible, the model may be deep and complex in order to fit the dataset. Therefore, overfitting problems may exist in a single deep neural network model when the number of weights in the deep neural network model is large. Besides, the learning rate of stochastic gradient descent (SGD) treats all parameters equally, resulting in local optimal solution. In this paper, the Pearson correlation coefficient is used to analyze the inherent correlation of PM2.5 and other auxiliary data such as meteorological data, season data, and time stamp data which are applied to cluster for enhancing the performance. Extracted features are helpful to build a deep ensemble network (EN) model which combines the recurrent neural network (RNN), long short-term memory (LSTM) network, and gated recurrent unit (GRU) network to predict the PM2.5 concentration of the next hour. The weights of the submodel change with the accuracy of them in the validation set, so the ensemble has generalization ability. The adaptive moment estimation (Adam) an algorithm for stochastic optimization is used to optimize the weights instead of SGD. In order to compare the overall performance of different algorithms, the mean absolute error (MAE) and mean absolute percentage error (MAPE) are used as accuracy metrics in the experiments of this study. The experiment results show that the proposed method achieves an accuracy rate (i.e., MAE = 6.19 and MAPE = 16.20%) and outperforms the comparative models.

1. Introduction

In recent years, the rapid development of the industry is accompanied by air pollution which causes the death of 7 million people every year and attracts great attention worldwide [1, 2]. Among these air pollutants, PM2.5 can traverse the nasal passages during inhalation and reach the throat and even the lungs [3] and brings about a great threat to the human body. In 2018, Heft-Neal et al. [4] suggested that PM2.5 concentration above minimum exposure levels was responsible for 22% of infant deaths in the 30 studied countries and led to 449,000 additional deaths of infants in 2015, an estimate that is more than three times higher than existing estimates that attribute the death of infants to poor air quality for these countries. Therefore, air control and prevention of air pollution have become significant issues. In order to achieve this goal, obtaining real-time air pollution concentration is necessary [5]. Moreover, sensors have been used in a wide range of applications [6, 7], which collect extensive air quality data. For the increased attention of air pollution, many researchers take a significant focus on air pollution and there are many relevant research studies about air pollution. The main machine learning methods applied to air pollution are as follows: artificial neural network (ANN), ensemble learning, support vector machine (SVM), and other hybrid models [8]. However, these existing prediction machine learning methods of air quality lack analyzing the reasons for the change of air pollution concentration because most of the prediction methods take more focus on the model selection and ignore the reasons for changing. Furthermore, since the framework of recent deep learning is very flexible, the model may be deep and complex in order to fit

the dataset. Therefore, overfitting problems may exist in a single deep neural network model when the number of weights in the deep neural network model is large. This paper analyzes the inherent relation of PM2.5 with other meteorological data (i.e., dew point, humidity, atmospheric pressure, temperature, wind direction, accumulated wind speed, precipitation, and accumulated precipitation), season data, and time stamp data. By analyzing the correlation between PM2.5 and other auxiliary data (an hour before), extracted air pollution characteristics are used to cluster the dataset and build a deep ensemble network (EN) model to predict PM2.5 concentration. The input of the model is the PM2.5 concentration and auxiliary data of the previous eight hours while the output is the PM2.5 concentration of the next hour. The adaptive moment estimation (Adam) algorithm is used to replace stochastic gradient descent (SGD) to update weights to get higher accuracy. For the validation of the proposed method, hourly PM2.5 concentration and meteorological data at 3 stations in Shanghai from 01 January 2010 to 31 December 2015 are collected. The mean absolute error (MAE) and mean absolute percentage error (MAPE) are used as accuracy metrics to compare the overall performance of each algorithm.

The contributions of this study are summarized as follows:

(i) This study proposes an ensemble model based on RNN, LSTM, and GRU to predict the PM2.5 concentration of the next hour

(ii) This study proposes a cluster method based on wind direction to improve prediction performance

(iii) Wind direction has been proved to be related to PM2.5 concentration because the wind can carry or take away PM2.5

The remainder of this paper is organized as follows. The literature reviews on air quality prediction in Section 2. Section 3 analyzes the inherent correlation between PM2.5 concentration and other auxiliary data and shows the data preprocess. Section 4 introduces the process of the Adam algorithm and the proposed EN. The experimental results are shown in Section 5. Section 6 is the conclusion including contributions and future work.

2. Literature Reviews

In this section, the disadvantages and advantages of existing machine learning for air quality prediction models are discussed.

2.1. Traditional Machine Learning Methods and Neural Networks with Simple Structure. Traditional machine learning methods and neural networks with simple structure were applied in PM2.5 prediction. In 2015, Lary et al. [9] proposed a model based on machine learning to estimate PM2.5 concentration. They collected the hourly PM2.5 data from 55 countries to verify the performance of the proposed model. Though the method got certain results, it could not predict future PM2.5 concentration. Hooyberghs et al. [10] proposed a method that combines ANN and big data to predict air quality. For the validation of the proposed model, the pollutant data, traffic data, and weather data were selected by smartphone sensors. The results showed that the ANN model is skilled in air pollution prediction. Moreover, some variations of ANN were proposed to predict air quality concentration. For example, the recurrent neural network (RNN) was used by Prakash et al. in 2011. This model was used to forecast 1 h ahead concentration and daily mean and daily maximum concentration of various pollutants, and the experimental results demonstrate the practicability of the method [11]. Besides, a hybrid model is also one of the variations of ANN. In 2011, Feng et al. [12] proposed a hybrid model that combines SVM with a back-propagation neural network (BPNN) to forecast ozone concentration. SVM was used to classify the data into its corresponding categories, and a genetic algorithm- (GA-) optimized BPNN was employed to build the prediction model. They collected the data including temperature, humidity, wind speed, and ultraviolet radiation from March 2009 to July 2009 to validate the accuracy of the proposed method, and the results showed that the model had a great prediction capability which could be used to predict the ozone concentration of Beijing. Considering the long-term dependencies and spatial correlations of air pollution, Li et al. [13] proposed an extended model of the long short-term memory (LSTM) network to extract the inherent features of air pollution and predict air quality. For the validation of the method, some models including the spatiotemporal deep learning (STDL) model, the time delay neural network (TDNN) model, the autoregressive moving average (ARMA) model, the support vector regression (SVR) model, and the traditional LSTM network [14] were used as the comparison algorithm and the results demonstrated the superiority of the proposed method.

2.2. Complex Deep Neural Networks. In recent years, deep learning has promoted the development of PM2.5 prediction. More and more complex deep networks are applied in this field to obtain better fitting results. In 2018, Huang and Kuo [15] analyzed the source pie of PM2.5 and proposed a deep neural network model that combines the convolution neural network (CNN) and LSTM network in 2018. CNN is a weight sharing network, which is good at capturing local features [16, 17]. This innovation of this method was introducing the convolution layer to extract spatial dependencies of PM2.5 and long short-term memory to extract temporal dependencies. The experimental results were compared with SVM, random forest (RD), decision tree (DT), NN, and LSTM algorithms and showed that PM2.5 concentration prediction models based on deep neural networks (e.g., NN, RNN, CNN, and LSTM) are better than the models based on traditional machine learning methods (e.g., SVM, RD, and DT). Considering the spatiotemporal dependence of PM2.5, Xie et al. [18] proposed a CGRU model based on CNN and gated recurrent unit (GRU) to predict PM2.5 concentration in the next six hours in 2019. CNN is used to extract spatial correlation features, and GRU further extracts long-term correlation features. Experimental results showed

that the proposed model was better than the traditional time series models (including LSTM, GRU, and ARIMA). GRU is a variant of LSTM, and the network structure is simpler than LSTM [19, 20]. In 2019, Tao et al. [21] proposed the CBGRU model based on 1D convnets and bidirectional GRU. On the basis of the bidirectional gated recurrent unit (BGRU), this method added the convolution layer and pool layer which can extract the PM2.5 features more easily. In 2020, Xayasouk et al. [22] developed two models including the LSTM model and the deep autoencoder (DAE) model to predict the particle concentration in the next hour. The experimental results showed that LSTM was better than DAE in predicting the particle concentration. In 2020, Kaya and Oguducu [23] proposed a new air quality prediction model based on deep learning, namely, deep flexible sequence. They used hourly data from Istanbul, Turkey, from 2014 to 2018 to predict air pollution before 4, 12, and 24 hours. This model is a hybrid and flexible deep model, which includes long short-term memory and convolutional neural network (CLSTM). On this basis, Li et al. [24] developed a deep CNN-LSTM method based on attention (ACLSTM), which includes the one-dimensional CNN, LSTM network, and attention network for urban PM2.5 concentration prediction. However, the attention layer is applied between the hidden layer and the output layer, which cannot explain the correlation between predictors and pollutants. Considering there are rare monitoring stations in a vast area, Ma et al. [25] proposed a deep spatiotemporal prediction method based on bidirectional LSTM and inverse distance weighting which can predict the PM2.5 concentration in the area without monitoring stations. Qi et al. [26] proposed a deep air learning method which provided novel ideas of interpolation, prediction, and feature analysis.

2.3. Ensemble Neural Networks. However, complex and deep network structure causes the decline of generalization ability, which leads to bad performance in other datasets. Hornik et al. [27] had proved that a single-layer ANN could approach the function with any complexity. However, how to make an appropriate network configuration was an NP-hard problem which influenced the generalization ability of the network. For solving the problem, Hansen and Salamon [28] proposed an ensemble neural network to provide a simple and feasible method. By this method, each of NN in the system was trained separately and the predictions of NN were synthesized as the final results.

3. Data Analysis

This paper collected hourly PM2.5 concentration and meteorological data at 3 stations in Shanghai from 01 January 2010 and 31 December 2015 from the UCI database [29]. The Pearson correlation coefficient is used to analyze the inherent correlation between PM2.5 and other auxiliary data of an hour before. The extracted features are used to choose the appropriate activation functions and train the EN which combined the RNN, LSTM, and GRU network to predict PM2.5 concentration. Before training the models, data preprocessing is necessary. This section analyzes the inherent correlation of PM2.5 and other auxiliary data in Section 3.1, and the data preprocessing is illustrated in Section 3.2.

3.1. Analyzing the Inherent Correlation of PM2.5 and Other Auxiliary Data. First of all, this study analyzes the spatial-temporal characteristics of three monitoring stations in Shanghai. They are located in Jingan, American consulate, and Xuhui separately as is shown in Figure 1. The autocorrelation function which is shown as Equation (1) is applied to measure the temporal correlation of each station, and more details are given in Reference [30]. In Table 1, the correlation values of three stations between PM2.5 concentration of an hour before and PM2.5 concentration of an hour later are above 0.95, which demonstrates that PM2.5 has a strong correlation in time.

$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}, \tag{1}$$

where X_t denotes the PM2.5 concentration of each hour, μ is the expectation of X_t, τ is the time delay, σ is the standard deviation, and E is the expectation function. Existing studies have proved that meteorological factors play a significant role in air pollutant concentration [31, 32]. Therefore, it is necessary to find out the relationship between meteorological factors and PM2.5 concentration. The Pearson correlation coefficients of PM2.5 and other auxiliary data are shown in Table 1. The PM2.5 concentration data is the next hour data, and the auxiliary data is the before hour data. DEWP stands for dew point, HUMI stands for humidity, PRES stands for atmospheric pressure, TEMP stands for temperature, CV stands for no wind, NE stands for northeast wind, SE stands for southeast wind, SW stands for southwest wind, NW stands for northwest wind, lws stands for accumulated wind speed, and Iprec stands for accumulated precipitation. As is shown in Table 1, except for HUMI, precipitation, Iprec, and spring, whose values are below 0.10, other auxiliary data values are above 0.10. In the respect of season, summer and autumn show a negative correlation with PM2.5 data concentration while winter has a positive correlation with PM2.5 data concentration. In the respect of wind direction, there is a negative correlation between PM2.5 data concentration and east wind while west wind has a positive correlation with PM2.5 data concentration. The Pearson correlation coefficient functions are shown as Equation (2), and more details are given in Reference [30].

$$R(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{2}$$

where σ_X is the standard deviation of X and σ_Y is the standard deviation of Y. μ_X is the expectation of X, and μ_Y is the expectation of Y.

For intuitively observing the correlation between PM2.5 concentration and meteorological data, the violin plots are shown in Figure 2. The abscissa is the interval of meteorological data, and the ordinate stands for the PM2.5 concentration in the interval (e.g., the temperature value ranges from

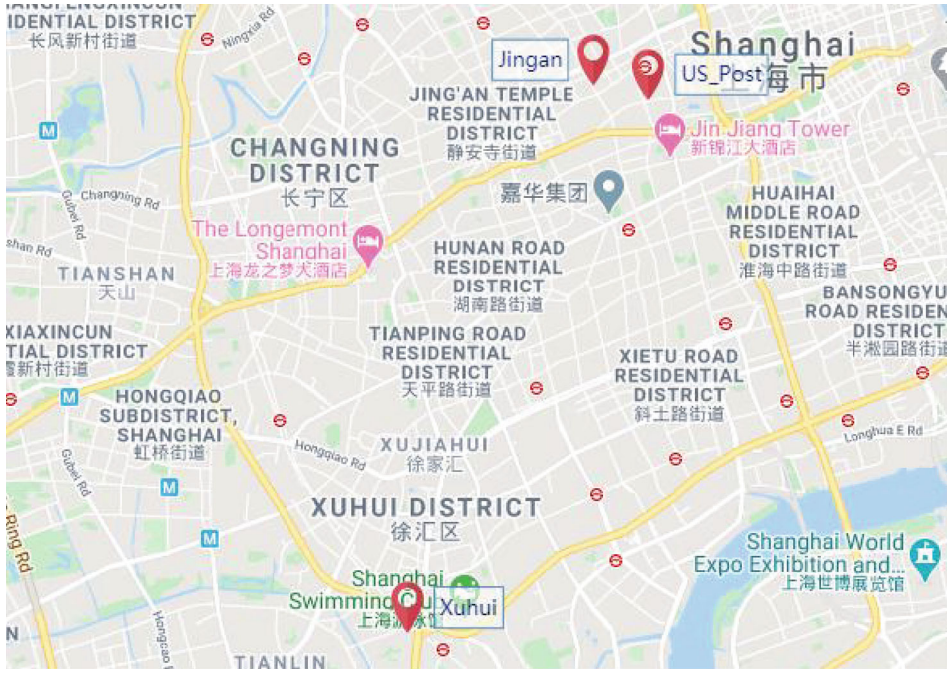


Figure 1: Distribution of PM2.5 monitoring stations in Shanghai (Jingan is located at 31°13′N, 121°26′E. Xuhui and US_Post are located at 31°10′N, 121°25′E).

-3 to 41, which is divided into 5 intervals as abscissa). The wider the image is, the more the number of data is. The division of interval varies with meteorological data. For the space limitation, this part only lists the violin plots of Jingan. Based on the correlation coefficient between wind directions and PM2.5 concentration, the dataset is divided into two parts: (1) west wind and no wind and (2) east wind. Two datasets are used to train and test the NN model. Six groups of controlled experiments are set to demonstrate the performance of the proposed method. Each group generates two models: (1) the NN model with the full dataset and (2) the NN model with the cluster method. In these groups, the number of dense layers of NN is 2 and the number of neurons is set from a candidate set of {5, 10, 15, 20, 25, 30}. Table 2 shows the experimental results for each test running. NN+CLU denotes NN based on the cluster method. In all groups, NN+CLU has the best performance (i.e., NN+CLU: 6.56 of MAE and 17.77% of MAPE in Group 1; NN+CLU: 6.87 of MAE and 19.65% of MAPE in Group 2; NN+CLU: 7.12 of MAE and 19.78% of MAPE in Group 3; NN+CLU: 6.92 of MAE and 19.11% of MAPE in Group 4; NN+CLU: 6.83 of MAE and 19.00% of MAPE in Group 5; and NN+CLU: 7.15 of MAE and 20.14% of MAPE in Group 6) which shows that the accuracy can be improved by the cluster method based on wind directions. When it is west direction, the PM2.5 concentration is higher, for Shanghai is located in the eastern coastal area of China, whose west is the inland. The west wind carries inland pollution, and no wind is not conducive to air circulation. On the contrary, the PM2.5 concentration is smaller, for the east wind carries the air from the ocean. The results illustrate that the cluster method based on wind directions can extract the data features effectively and get better predicted results.

3.2. Data Preprocessing. Hourly PM2.5 concentration, season data, and meteorological data at 3 stations in Shanghai from 01 January 2010 to 31 December 2015 are collected to test the performance of the proposed hourly PM2.5 concentration and meteorological data at 3 stations in Shanghai from 01 January 2010 to 31 December 2015 which are collected to test the performance of the proposed method. In the data preprocessing stage, the records with abnormal values or missing values are deleted firstly. Secondly, the wind direction data are changed into binary codes to enhance the prediction performance. Wind direction index data have 5 unique categorical values, and each wind direction index is transferred to a 5-dimensional vector (e.g., northwest is assigned as [0, 0, 0, 0, 1]). Similarly, season index data have 4 unique categorical values, and each season index is transferred to a 4-dimensional vector (e.g., spring is assigned as [1, 0, 0, 0]). Thirdly, data is processed by the min-max normalization method and compressed from 0 to 1 for a better training effect. The formula is shown as Equation (3), and more details are given in Reference [33].

$$z_i = \frac{x_i - \min\limits_{1\le n\le N}(x_n)}{\max\limits_{1\le n\le N}(x_n) - \min\limits_{1\le n\le N}(x_n)}, \tag{3}$$

where n denotes n-th records and N is the number of records. z is the normalized data ranging from 0 to 1.

4. The Proposed Ensemble Network Model for Prediction with Adam

The proposed EN model can be divided into three submodels: RNN, LSTM, and GRU. Three network models are used

Table 1: The correlation coefficient of the before hour auxiliary data and the next hour of PM2.5 concentration.

	Jingan	US_Post	Xuhui	DEWP	HUMI	PRES	TEMP	lws	Precipitation	Iprec	CV	NE	SE	SW	NW	Spring	Summer	Autumn	Winter
Jingan	0.97	0.96	0.94	-0.22	-0.06	0.14	-0.21	-0.23	-0.07	-0.09	0.12	-0.23	-0.14	0.15	0.27	0.02	-0.15	-0.15	0.29
US_Post	0.96	0.96	0.91	-0.33	-0.08	0.25	-0.33	-0.22	-0.08	-0.09	0.11	-0.19	-0.16	0.12	0.28	0.03	-0.25	-0.15	0.36
Xuhui	0.94	0.91	0.97	-0.25	-0.08	0.17	-0.24	-0.23	-0.08	-0.09	0.12	-0.23	-0.13	0.15	0.26	0.00	-0.16	-0.14	0.31
DEWP	-0.22	-0.33	-0.25	0.99	0.42	-0.85	0.87	-0.01	0.09	0.07	-0.06	-0.05	0.20	0.09	-0.22	-0.12	0.62	0.15	-0.64
HUMI	-0.06	-0.08	-0.08	0.42	0.95	-0.21	-0.05	0.03	0.15	0.15	0.04	0.08	0.07	-0.15	-0.07	-0.09	0.13	0.01	-0.05
PRES	0.14	0.25	0.17	-0.85	-0.21	1.00	-0.84	0.00	-0.09	-0.07	0.07	0.17	-0.20	-0.18	0.15	-0.07	-0.67	0.11	0.61
TEMP	-0.21	-0.33	-0.24	0.87	-0.05	-0.84	0.99	-0.03	0.02	0.01	-0.08	-0.11	0.18	0.18	-0.20	-0.07	0.61	0.15	-0.68
lws	-0.23	-0.22	-0.23	-0.01	0.03	0.00	-0.03	0.95	0.03	0.06	-0.13	0.12	0.04	-0.15	-0.01	0.03	-0.04	0.03	-0.02
Precipitation	-0.07	-0.08	-0.08	0.09	0.15	-0.09	0.02	0.03	0.38	0.21	-0.02	0.03	0.00	-0.02	-0.02	0.00	0.05	-0.01	-0.03
Iprec	-0.09	-0.09	-0.09	0.07	0.15	-0.07	0.01	0.06	0.21	0.91	-0.02	0.07	-0.03	-0.03	-0.02	0.02	0.02	0.02	-0.02
CV	0.12	0.11	0.12	-0.06	0.04	0.07	-0.08	-0.13	-0.02	-0.02	0.31	-0.07	-0.06	0.05	-0.03	-0.08	-0.09	0.15	0.01
NE	-0.23	-0.19	-0.23	-0.05	0.08	0.17	-0.11	0.12	0.03	0.07	-0.07	0.76	-0.38	-0.27	-0.23	0.15	0.09	-0.11	-0.15
SE	-0.14	-0.16	-0.13	0.20	0.07	-0.20	0.18	0.04	0.00	-0.03	-0.06	-0.38	0.75	-0.13	-0.30	0.02	0.14	-0.11	-0.05
SW	0.15	0.12	0.15	0.09	-0.15	-0.18	0.18	-0.15	-0.02	-0.03	0.05	-0.27	-0.13	0.62	-0.10	-0.09	-0.11	0.03	0.17
NW	0.27	0.28	0.26	-0.22	-0.07	0.15	-0.20	-0.01	-0.02	-0.02	-0.03	-0.23	-0.30	-0.10	0.75	1.00	-0.33	-0.36	-0.35

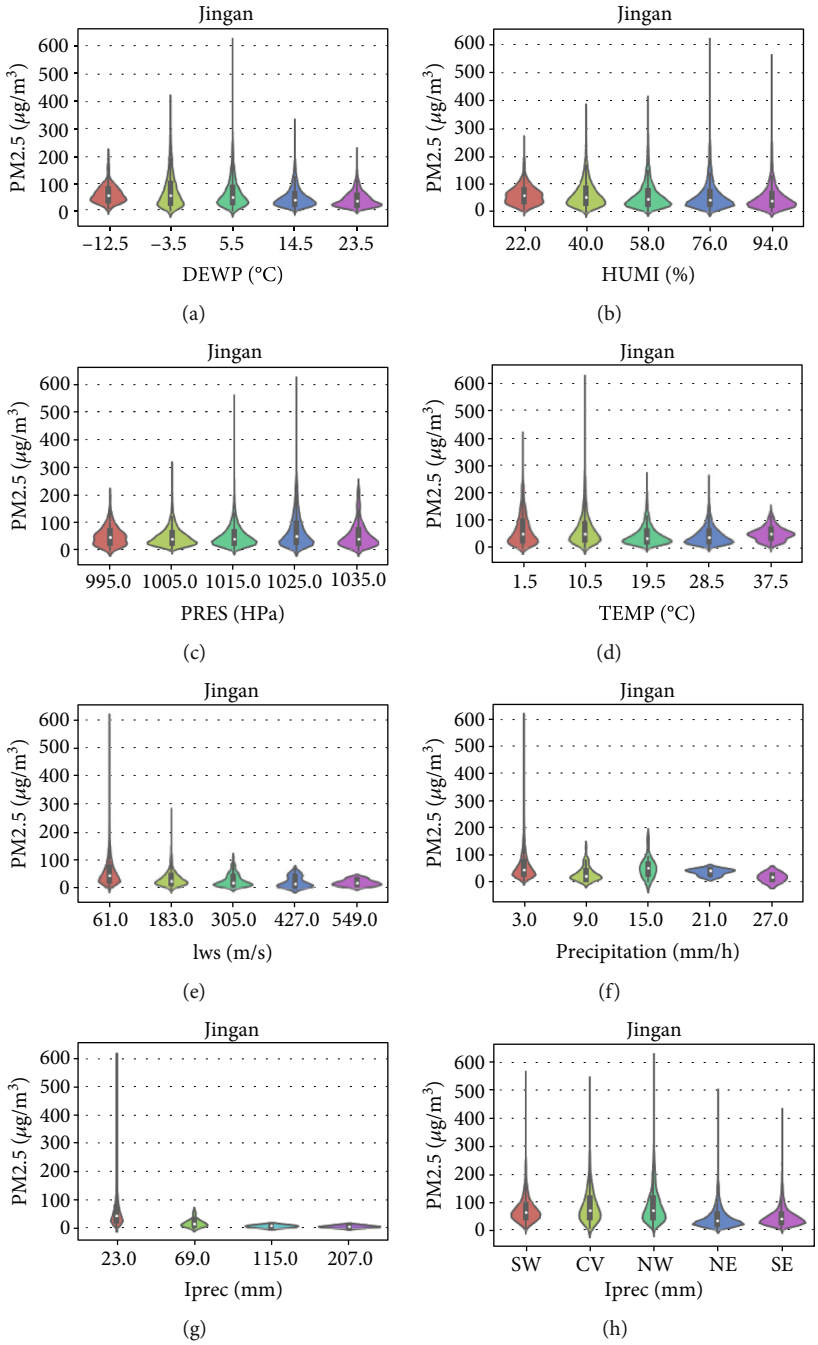


Figure 2: The violin plots between PM2.5 and meteorological data.

to predict PM2.5 concentration, respectively, and the results of them are combined by the weighted average method. The weights of EN are optimized by the Adam algorithm. Section 4.1 illustrates the EN, and Section 4.2 introduces the algorithm process of Adam.

4.1. Ensemble Network. As shown in Figure 3, the proposed EN model consisting of the RNN model, LSTM model, and GRU model can be applied to predict PM2.5 concentration. The establishment of the proposed model can be divided into two sections: training independent network model and combining the results of every network model. This section introduces the training stage of the NN, RNN, LSTM, and GRU models in Sections 4.1.1 and 4.1.2. The combining stage is illustrated in Section 4.1.3.

4.1.1. Neural Network. In this stage, historical and auxiliary data are used to train each network model in the EN model separately. The NN model is employed to analyze the interaction effects of input parameters and get the prediction. It is divided into three layers: input layer, hidden layer, and output layer. Besides, the rectified linear unit (RELU) function is applied as an activation function which is added behind the output layer to produce a nonlinear prediction. The

Table 2: The accuracy comparisons between NN and NN+CLU.

	Model	The number of neurons of dense layers	MAE	MAPE
Group 1	NN	5	6.99	19.59%
	NN +CLU		6.56	17.77%
Group 2	NN	10	6.89	19.67%
	NN +CLU		6.87	19.65%
Group 3	NN	15	7.13	19.89%
	NN +CLU		7.12	19.78%
Group 4	NN	20	6.95	19.19%
	NN +CLU		6.92	19.11%
Group 5	NN	25	7.00	19.96%
	NN +CLU		6.83	19.00%
Group 6	NN	30	7.21	20.29%
	NN +CLU		7.15	20.14%

Adam an algorithm for stochastic optimization is used to optimize the weights instead of SGD.

For the training of NNs, the PM2.5 concentration data and auxiliary data of the previous hour in three stations are used as the inputs of NNs while the PM2.5 concentration data of the next hour in three stations are used as the outputs of NNs. The full connected layers are used as the hidden layer to analyze the inherent correlation of parameters. The historical data are applied to train the models, and the weights of models are optimized by the Adam algorithm.

4.1.2. Recurrent Neural Network. Besides NN models, other deep learning methods are applied to prediction problems (e.g., RNN models, LSTM models). The inputs of the RNN, LSTM, and GRU models are different from NN models for each neuron of the NN's input layers is a single sequence element while each neuron of the input layer in RNN, LSTM, and GRU is a vector which is encoded by the past sequence elements. As has been said before, PM2.5 concentration data have a strong correlation in time. RNN, LSTM, and GRU are applied to predict PM2.5 concentration for they are experts in dealing with time series problems compared with NN models. LSTM an extended model of RNN differs from RNN in learning long-time dependence for there is a phenomenon of gradient disappearance in RNN. GRU an extended model of LSTM differs from LSTM in internal structure for LSTM has three gates and GRU has only two. In the training stage of the RNN, LSTM, and GRU models, eight hours of PM2.5 concentration data along with meteorological data is regarded as input which is different from the NN model.

4.1.3. Combining State of the Ensemble Network. In Figure 3, each submodel in EN can predict the PM2.5 concentration independently and all the results will be integrated to produce the final prediction results. In this research, the number of hidden layer neurons ranged from 5 to 30 increasing by 5 at a time. The weighted average method is applied to integrate all the prediction results of submodels. In order to obtain the weight of each submodel, 10% of the dataset is selected as the validation set. The accuracy of each submodel is applied in Softmax to get weight. The Softmax formula is shown in Equation (4), and more details are given in [34].

$$w_i = \frac{e^{z_i}}{\sum_i^n e^{z_i}}, \tag{4}$$

where $z_1, z_2, \cdots, z_i$ denote the accuracy of submodels on the validation set, n is the number of submodels, and $w_1, w_2, \cdots, w_i$ are the weight of submodels. e denotes the natural exponential. The final result of the proposed model can be computed as

$$\text{accuracy} = \sum_i^n w_i \cdot z_i. \tag{5}$$

This section sets six groups of controlled experiments. Each group generates seven models consisting of NN, RNN, LSTM, EN, EN1, EN2, EN3, EN4, and EN5, where EN denotes a combination of RNN, LSTM, and GRU; EN1 denotes a combination of NN and RNN; EN2 denotes a combination of NN and LSTM; EN3 denotes a combination of RNN and LSTM; EN4 denotes a combination of NN, RNN, and LSTM; EN5 denotes a combination of NN, RNN, LSTM, and GRU. In these groups, the number of dense layers of NN, RNN, LSTM, and GRU is 1 and the number of neurons is generated from a candidate set of {5, 10, 15, 20, 25, 30}. For the comparisons of the overall performance of different optimized algorithms, MAE and MAPE are used as accuracy metrics for the experiments.

Table 3 shows the experimental results for each test running. In all groups, EN (including EN, EN1, EN2, EN3, EN4, and EN5) has the best performance (i.e., EN1: 6.51 of MAE and 19.35% of MAPE in Group 1; EN: 6.35 of MAE and 17.36% of MAPE in Group 2; EN: 6.19 of MAE and 16.20% of MAPE in Group 3; EN: 6.20 of MAE and 16.28% of MAPE in Group 4; EN2: 6.71 of MAE and 19.56% of MAPE in Group 5; and EN3: 6.47 of MAE and 17.80% of MAPE in Group 6). The results illustrate that the combination of different models is effective to predict PM2.5 concentration and decreases the error of overfitting.

4.2. Adam Optimization. This paper adopted Adam [35] an optimization algorithm to replace the traditional SGD. It can update the weights of the neural network iteratively based on the dataset. The Adam optimization algorithm is an extension of the SGD algorithm, which is widely used in deep learning applications recently. The Adam algorithm is different from traditional SGD for the latter keeps a single learning rate to update all weights and the learning rate does not change in the training process while the former designs

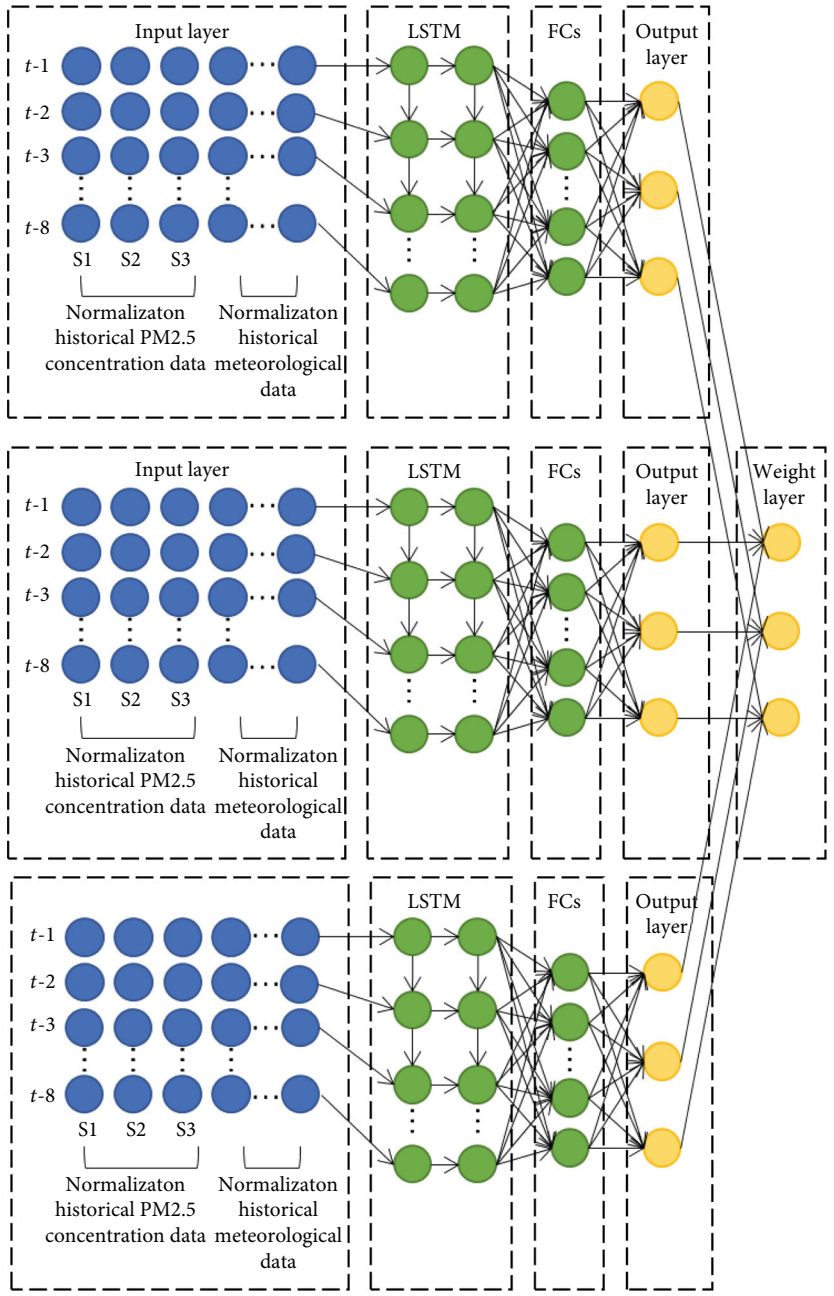


Figure 3: Structure of the EN model for PM2.5 forecasting.

an independent adaptive learning rate for different parameters. The Adam algorithm will be introduced in detail next.

Assume that f is the objective function and θ are the parameters which require to be optimized. g_t stands for the gradient which can be expressed as Equation (6), and more details of formulas of this section are given in [35].

$$g_t = \nabla_\theta f_t(\theta_{t-1}), \tag{6}$$

where the $f_1(\theta), f_2(\theta), \cdots, f_t(\theta)$ stand for the function values of time step 1 to t. m_t and v_t stand for the exponential moving averages of the gradient (i.e., biased first moment estimate) and the squared gradient (i.e., biased second raw moment estimate) which are employed to update weights separately and their formulas are shown as

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t, \tag{7}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot {g_t}^2, \tag{8}$$

where β_1 and β_2 ranging from 0 to 1 control the exponential decay rates for the moment estimates.

In the beginning, m_t and v_t are near 0, for they are set as 0 and the decay rate is close to 1. For the sake of counteracting the bias at the beginning, bias-corrected estimates m_t and v_t

Table 3: The accuracy comparisons between NN, RNN, LSTM, and EN with different numbers of neurons.

	Model	The number of neurons of dense layers	MAE	MAPE
Group 1	NN	5	6.99	19.59%
	RNN	5	7.23	20.17%
	LSTM	5	6.63	20.01%
	GRU	5	6.73	20.23%
	EN1 (NN+RNN)	{5, 5}	7.19	20.31%
	EN2 (NN+LSTM)	{5, 5}	6.62	19.65%
	EN3 (RNN+LSTM)	{5, 5}	6.59	19.35%
	EN4 (NN+RNN+LSTM)	{5, 5, 5}	6.73	19.61%
	EN5 (NN+RNN+LSTM+GRU)	{5, 5, 5, 5}	6.72	19.60%
	EN (RNN+LSTM+GRU)	{5, 5, 5}	6.60	19.99%
Group 2	NN	10	6.89	19.17%
	RNN	10	6.57	18.70%
	LSTM	10	8.27	25.43%
	GRU	10	7.23	19.88%
	EN1 (NN+RNN)	{10, 10}	6.54	18.30%
	EN2 (NN+LSTM)	{10, 10}	7.02	20.76%
	EN3 (RNN+LSTM)	{10, 10}	6.84	20.68%
	EN4 (NN+RNN+LSTM)	{10, 10, 10}	6.65	19.49%
	EN5 (NN+RNN+LSTM+GRU)	{10, 10, 10, 10}	6.67	19.23%
	EN (RNN+LSTM+GRU)	{10, 10, 10}	6.35	17.36%
Group 3	NN	15	7.11	19.89%
	RNN	15	6.57	16.77%
	LSTM	15	6.45	17.34%
	GRU	15	6.56	19.84%
	EN1 (NN+RNN)	{15, 15}	6.42	17.68%
	EN2 (NN+LSTM)	{15, 15}	6.49	17.06%
	EN3 (RNN+LSTM)	{15, 15}	6.22	16.25%
	EN4 (NN+RNN+LSTM)	{15, 15, 15}	6.26	16.64%
	EN5 (NN+RNN+LSTM+GRU)	{15, 15, 15, 15}	6.25	16.56%
	EN (RNN+LSTM+GRU)	{15, 15, 15}	6.19	16.20%
Group 4	NN	20	6.95	19.19%
	RNN	20	7.23	18.49%
	LSTM	20	6.48	16.79%
	GRU	20	6.60	17.25%
	EN1 (NN+RNN)	{20, 20}	6.57	17.39%
	EN2 (NN+LSTM)	{20, 20}	6.20	16.42%
	EN3 (RNN+LSTM)	{20, 20}	6.46	16.57%
	EN4 (NN+RNN+LSTM)	{20, 20, 20}	6.24	16.28%
	EN5 (NN+RNN+LSTM+GRU)	{20, 20, 20, 20}	6.20	16.56%
	EN (RNN+LSTM+GRU)	{20, 20, 20}	6.28	16.46%
Group 5	NN	25	7.00	19.96%
	RNN	25	7.79	23.55%
	LSTM	25	7.14	21.04%
	GRU	25	7.08	19.39%
	EN1 (NN+RNN)	{25, 25}	7.10	20.87%
	EN2 (NN+LSTM)	{25, 25}	6.71	19.56%
	EN3 (RNN+LSTM)	{25, 25}	7.13	21.57%
	EN4 (NN+RNN+LSTM)	{25, 25, 25}	6.86	20.37%

Table 3: Continued.

	Model	The number of neurons of dense layers	MAE	MAPE
Group 6	EN5 (NN+RNN+LSTM+GRU)	{25, 25, 25, 25}	6.88	20.22%
	EN (RNN+LSTM+GRU)	{25, 25, 25}	6.80	19.78%
	NN	30	7.21	20.29%
	RNN	30	7.02	19.74%
	LSTM	30	6.91	18.65%
	GRU	30	7.02	19.56%
	EN1 (NN+RNN)	{30, 30}	6.90	19.41%
	EN2 (NN+LSTM)	{30, 30}	6.66	18.51%
	EN3 (RNN+LSTM)	{30, 30}	6.57	18.34%
	EN4 (NN+RNN+LSTM)	{30, 30, 30}	6.60	18.48%
	EN5 (NN+RNN+LSTM+GRU)	{30, 30, 30, 30}	6.56	18.23%
	EN (RNN+LSTM+GRU)	{30, 30, 30}	6.47	17.80%

are introduced. The calculation formulas are shown as

$$m_t = \frac{m_t}{1 - \beta_1{}^t}, \tag{9}$$

$$v_t = \frac{v_t}{1 - \beta_2{}^t}. \tag{10}$$

The final updated formula of parameters is shown as

$$\theta_t = \theta_{t-1} - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}, , \tag{11}$$

where α denotes the stepsize (i.e., learning rate) and is initialized to 0.001 (i.e., default value).

This section sets four groups of controlled experiments, and NN, RNN, LSTM, and GRU are applied to predict the PM2.5 concentration by using SGD, Adam, and Nadam separately. All the number of neurons of dense layers in these models is set to 15, and the performance of each algorithm in different networks is shown in Table 4. Compared with Adam and Nadam, SGD has the worst performance because it has a fixed learning rate leading to find global optimum difficultly. Except for the MAE of NN, Adam has better performance than Nadam [36] which proves that the Adam algorithm is effective in optimizing the PM2.5 predicting model.

In order to determine the values of β_1 and β_2, this section sets three groups of controlled experiments and EN (RNN, LSTM, and GRU) is employed to predict PM2.5 concentration with different β_1 and β_2. All the number of neurons of dense layers in these models is set to 15; the performance of different values of β_1 and β_2 is shown in Table 5. The default values of β_1 and β_2 (i.e., $\beta_1 = 0.9$ and $\beta_2 = 0.999$) get the best performance comparing with other values (i.e., $\beta_1 = 0.8$, $\beta_2 = 0.888$ and $\beta_1 = 0.7$, $\beta_2 = 0.777$).

5. Experimental Results and Analysis

This section uses hourly PM2.5 concentration and meteorological data at 3 stations in Shanghai from 01 January 2010 to 31 December 2015 to evaluate the proposed model. All the models including NN, RNN, LSTM, GRU, BGRU, CGRU, CBGRU, CLSTM, ACLSTM, and EN are trained on the Keras framework with TensorFlow backend. The learning rate is set to 0.001, and the epochs are set to 200. RELU is applied as an activation function for each layer of the network, and Adam is used as the optimized algorithm to optimize the weights. For the validation of the proposed method, MAE and MAPE are used as accuracy metrics to compare the overall performance of each model. MAE is an absolute value, and MAPE is a percentage, smaller values of which indicate better performance. Two metrics are given in Equations (12) and (13), respectively, and more details are given in Reference [30].

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|o_n - p_n|, \tag{12}$$

$$\text{MAPE} = \frac{1}{N}\sum_{n=1}^{N}\frac{|o_n - p_n|}{o_n}, \tag{13}$$

where o_n is the value of the n-th observed data and p_n denotes the predicted value of the n-th predicted data. The values of two metrics (MAE, MAPE) are calculated for proposed EN as well as for NN, RNN, LSTM, GRU, BGRU, CGRU, CBGRU, CLSTM, and ACLSTM. Table 6 illustrates the detailed MAE and MAPE values of each algorithm. In the ranking of MAE from high to low, we have CBGRU (8.58), CGRU (8.56), RNN (7.23), BGRU (7.22), NN (6.95), CLSTM (6.90), ACLSTM (6.86), GRU (6.56), LSTM (6.48), and EN (6.19). In the ranking of MAPE from high to low, we have CBGRU (24.73%), CGRU (20.50%), GRU (19.84%), NN (19.19%), BGRU (18.88%), RNN (18.49%), CLSTM (17.41%), LSTM (16.79%), ACLSTM (16.60%), and EN (16.20%). No matter MAE or MAPE, EN has the minimum values which confirms the advantages of EN. Compared with the traditional NN and RNN, LSTM and GRU have better performance in predicting PM2.5 which indicates that the PM2.5 concentration has a long-term dependence. As submodels of the ensemble model, NN, RNN, LSTM, and GRU

Table 4: The accuracy comparisons of different models with different optimized algorithms.

Model	The number of neurons of dense layers	Algorithm	MAE	MAPE
NN	15	SGD	7.13	22.38%
	15	Adam	7.11	19.89%
	15	Nadam	6.75	20.11%
RNN	15	SGD	12.33	32.45%
	15	Adam	6.57	16.77%
	15	Nadam	6.58	17.29%
LSTM	15	SGD	11.18	30.38%
	15	Adam	6.45	17.34%
	15	Nadam	6.67	17.47%
GRU	15	SGD	9.19	25.00%
	15	Adam	6.56	19.84%
	15	Nadam	6.70	19.98%

Table 5: The accuracy comparisons of different values of β_1 and β_1 in the EN model.

Values of β_1 and β_1	The number of neurons of dense layers	MAE	MAPE
$\beta_1 = 0.9$ $\beta_2 = 0.999$	15	6.19	16.20%
$\beta_1 = 0.8$ $\beta_2 = 0.888$	15	6.26	16.44%
$\beta_1 = 0.7$ $\beta_2 = 0.777$	15	6.20	16.35%

Table 6: The accuracy comparisons of different algorithms.

Model	MAE	MAPE
NN	6.95	19.19%
RNN	7.23	18.49%
LSTM	6.48	16.79%
GRU	6.56	19.84%
BGRU	7.22	18.88%
CGRU	8.56	20.50%
CBGRU	8.58	24.73%
CLSTM	6.90	17.41%
ACLSTM	6.86	16.60%
EN (RNN+LSTM+GRU)	6.19	16.20%

with fewer layers have achieved considerable results in predicting PM2.5 concentration. However, when the models become more complex and deeper, the prediction performance of the models does not necessarily develop in the direction of optimization. Although they can show the best performance in their respective datasets with deep and complex structures, they may lead to overfitting in other datasets, which decreases the generalization ability of the models. The ensemble network is a model that can adapt to different datasets for the ensemble model consists of three submodels and the weight of each model depends on the prediction accuracy of the model in the dataset employed, which ensures the stability of the model in different datasets. Figure 4 shows the comparison between observed values and predicted values of the EN model. From the overall trend, the prediction data can better fit the observation data. This also proves that the proposed model can effectively predict the PM2.5 concentration in the next hour.

6. Conclusions and Future Work

Because of the flexibility of the network framework, many complex deep learning networks have been developed for air quality prediction. As far as we know, there is no uniform dataset in current air quality prediction research. Researchers collected datasets from different regions to train the network. Although these complex deep networks can well fit the data they use, they lack generalization ability. Therefore, this paper proposes an EN model to predict air pollution concentration by historical PM2.5 concentration, meteorological, and time stamp data. Considering that the submodel including RNN, LSTM, and GRU has quite good performance, each submodel of the EN model is trained, respectively, to get the accuracy and obtain the final model by a weighted average method. The weights of submodels are flexible because they are obtained by the accuracy of the validation set so that they can perform stably in different datasets. In addition, the ensemble of different networks is less involved in this field As far as we know, Adam is adopted to optimize weights instead of SGD for it can adjust the learning rate adaptively to get an efficient training effect. A case study of the prediction of PM2.5 concentration in Shanghai of the People's Republic of China is given in this research, and the dataset is divided into three parts: training data, validation data, and testing data. Training data are used to train the submodels of EN separately, validation data are applied to obtain the weight of each submodel, and testing data are adopted to compute MAE and MAPE for performance evaluation. The experimental evaluation is performed for EN, as well as other algorithms including NN, RNN, LSTM, GRU, BGRU, CGRU, CBGRU, CLSTM, and ACLSTM. The experimental results demonstrate that the proposed method has the best performance which outperforms other algorithms. Several findings of this paper are as follows:

(i) Compared with the single model, the EN model has better generalization ability and predictive ability as validated by MAE and MAPE

(ii) Wind direction has a significant impact on PM2.5 concentration for wind can carry or take away PM2.5

(iii) Compared with SGD, the Adam algorithm avoid the local optimum effectively

For the extension of this study, the prediction performance can be enhanced by adding human activities because it is one of the main reasons for environmental deterioration

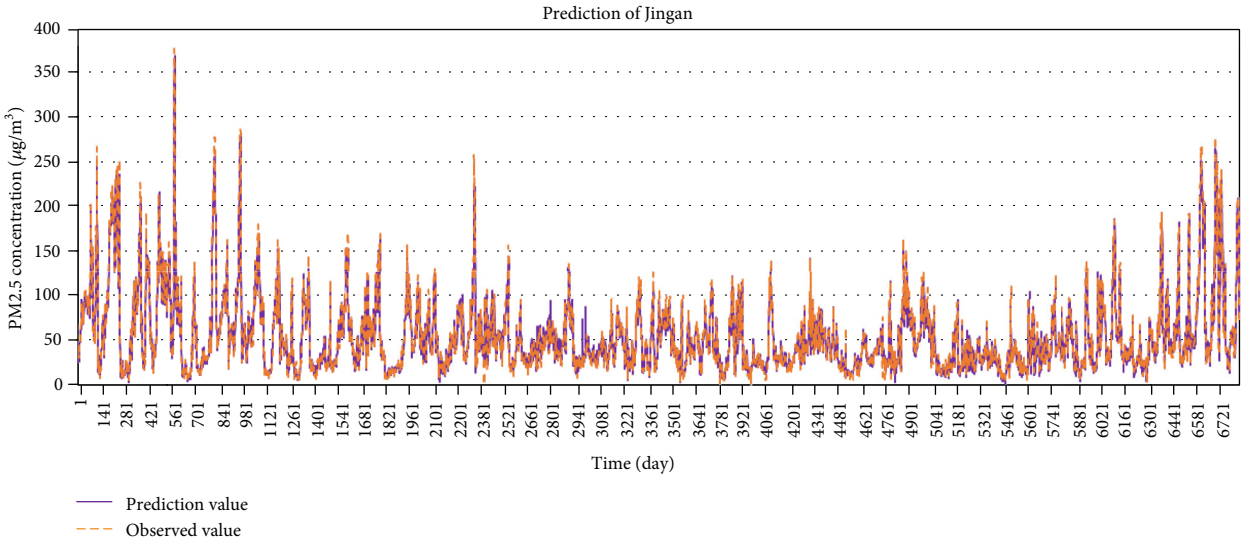


Figure 4: PM2.5 concentration forecasting results of the EN model.

especially in holidays. Furthermore, this paper found that wind direction has a significant influence on PM2.5 concentration because the wind will bring or take away PM2.5. However, it is uncertain that it can do in the areas with high mountains. Therefore, embedding the influence of topographical factors can become a research direction in the future. However, limited by lacking human activity and topographical data, this paper only analyzes the impact of meteorological data on PM2.5.

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.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Nos. 61906043, 61877010, 11501114, and 11901100), Fujian Natural Science Funds (No. 2019J01243), Funds of Department of Education, Fujian Province (No. JAT190026), and Fuzhou University (Nos. 510872/GXRC-20016, 510930/XRC-20060, 510730/XRC-18075, 510809/GXRC-19037, 510649/XRC-18049, and 510650/XRC-18050).

References

[1] F. Launay, "7 million deaths annually linked to air pollution," *Central European Journal of Public Health*, vol. 22, no. 1, pp. 53–59, 2014.

[2] A. Kurt and A. B. Oktay, "Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986–7992, 2010.

[3] D. W. Dockery, C. A. Pope, X. Xu et al., "An association between air pollution and mortality in six U.S. cities," *New England Journal of Medicine*, vol. 329, no. 24, pp. 1753–1759, 1993.

[4] S. Heft-Neal, J. Burney, E. Bendavid, and M. Burke, "Robust relationship between air quality and infant mortality in Africa," *Nature*, vol. 559, no. 7713, pp. 254–258, 2018.

[5] Y. Zheng, F. Liu, and H. P. Hsieh, "U-air: when urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1436–1444, Chicago, USA, 2013.

[6] X. S. Xue and J. F. Chen, "Optimizing sensor ontology alignment through compact co-firefly algorithm," *Sensors*, vol. 20, no. 7, article 2056, 2020.

[7] X. S. Xue and J. F. Chen, "Using compact evolutionary Tabu search algorithm for matching sensor ontologies," *Swarm and Evolutionary Computation*, vol. 48, pp. 25–30, 2019.

[8] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality Modelling: A Systematic Review," *Applied Sciences*, vol. 8, no. 12, article 2570, 2018.

[9] D. J. Lary, T. Lary, and B. Sattler, "Using machine learning to estimate global PM2.5 for environmental health studies," *Environmental Health Insights*, vol. 9, no. S1, article EHI. S15664, 2020.

[10] J. Hooyberghs, C. Mensink, G. Dumont, F. Fierens, and O. Brasseur, "A neural network forecast for daily average PM concentrations in Belgium," *Atmospheric Environment*, vol. 39, no. 18, pp. 3279–3289, 2005.

[11] U. K. Prakash, K. Kumar, and V. K. Jain, "A wavelet-based neural network model to predict ambient air pollutants' concentration," *Environmental Modeling & Assessment*, vol. 16, no. 5, pp. 503–517, 2011.

[12] Y. Feng, W. Zhang, D. Sun, and L. Zhang, "Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification," *Atmospheric Environment*, vol. 45, no. 11, pp. 1979–1985, 2011.

[13] X. Li, L. Peng, X. Yao et al., "Long short-term memory neural network for air pollutant concentration predictions: method

development and evaluation," *Environmental Pollution*, vol. 231, Part 1, pp. 997–1004, 2017.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] C. J. Huang and P. H. Kuo, "A deep CNN-LSTM Model for particulate matter (PM2.5) forecasting in smart cities," *Sensors*, vol. 18, no. 7, article 2220, 2018.

[16] I. S. Krizhevsky and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[17] Y. Kim, "Convolutional neural networks for sentence classification," 2014, https://arxiv.org/abs/1408.5882.

[18] H. F. Xie, L. Ji, Q. Wang, and Z. J. Jia, "Research of PM2.5 prediction system based on CNNs-GRU in Wuxi urban area," *Earth and Environmental Science*, vol. 300, 2020.

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, https://arxiv.org/abs/1412.3555.

[20] X. Luo, W. Zhou, W. Wang, Y. Zhu, and J. Deng, "Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data," *IEEE Access*, vol. 6, pp. 5705–5715, 2017.

[21] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU," *IEEE Access*, vol. 7, pp. 76690–76698, 2019.

[22] T. Xayasouk, H. Lee, and G. Lee, "Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models," *Sustainability*, vol. 12, no. 6, p. 2570, 2020.

[23] K. Kaya and S. G. Oguducu, "Deep flexible sequential (DFS) model for air pollution forecasting," *Scientific Reports*, vol. 10, no. 1, 2020.

[24] S. Z. Li, G. Xie, J. C. Ren, L. Guo, L. Y. Y. Yang, and X. Y. Xu, "Urban PM2.5 concentration prediction via attention-based CNN-LSTM," *Applied Sciences*, vol. 10, no. 6, 2020.

[25] J. Ma, Y. X. Ding, V. J. L. Gan, C. Q. Lin, and Z. W. Wan, "Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM," *IEEE Access*, vol. 7, pp. 107897–107907, 2019.

[26] Z. G. Qi, T. C. Wang, G. J. Song, W. S. Hu, X. Li, and Z. F. Zhang, "Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2285–2297, 2018.

[27] K. Hornik and M. Stinchcombe, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[28] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[29] S. X. Chen, "PM2.5 data of five Chinese cities data set," 2017, https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities.

[30] S. Kounev, K. D. Lange, and J. V. Kistowski, "Review of basic probability and statistics," in *Systems Benchmarking*, 2020.

[31] J. He, Y. Yu, Y. Xie et al., "Numerical model-based artificial neural network model and its application for quantifying impact factors of urban air quality," *Water, Air, & Soil Pollution*, vol. 227, no. 7, article 235, 2016.

[32] Y. Bai, Y. Li, X. Wang, J. Xie, and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," *Atmospheric Pollution Research*, vol. 7, no. 3, pp. 557–566, 2016.

[33] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[34] M. Y. Jiang, Y. C. Liang, X. Y. Feng et al., "Text classification based on deep belief network and Softmax regression," *Neural Computing & Applications*, vol. 29, no. 1, pp. 61–70, 2018.

[35] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[36] D. Timothy, "Incorporating Nesterov momentum into Adam," *Natural Hazards*, vol. 3, no. 2, pp. 437–453, 2016.