Hindawi
Complexity
Volume 2020, Article ID 6612043, 12 pages
https://doi.org/10.1155/2020/6612043

Research Article

Network Pseudohealth Information Recognition Model: An Integrated Architecture of Latent Dirichlet Allocation and Data Block Update

Jie Zhang, Pingping Sun, Feng Zhao, Qianru Guo, and Yue Zou

School of Economics and Management, Shandong University of Science and Technology, Qingdao 266590, China

Correspondence should be addressed to Feng Zhao; chinazhaof@163.com

Received 22 November 2020; Revised 5 December 2020; Accepted 10 December 2020; Published 21 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

The wanton dissemination of network pseudohealth information has brought great harm to people's health, life, and property. It is important to detect and identify network pseudohealth information. Based on this, this paper defines the concepts of pseudohealth information, data block, and data block integration, designs an architecture that combines the latent Dirichlet allocation (LDA) algorithm and data block update integration, and proposes the combination algorithm model. In addition, crawler technology is used to crawl the pseudohealth information transmitted on the Sina Weibo platform during the "epidemic situation" from February to March 2020 for the simulation test on the experimental case dataset. The research results show that (1) the LDA model can deeply mine the semantic information of network pseudohealth information, obtain the features of document-topic distribution, and classify and train topic features as input variables; (2) the dataset partitioning method can effectively block data according to the text attributes and class labels of network pseudohealth information and can accurately classify and integrate the block data through the data block reintegration method; and (3) considering that the combination model has certain limitations on the detection of network pseudohealth information, the support vector machine (SVM) model can extract the granularity content of data blocks in pseudohealth information in real time, thus greatly improving the recognition performance of the combination model.

1. Introduction

At present, pneumonia caused by the new coronavirus has been effectively controlled nationwide, but the panic and fear caused by it are making people nervous. People are attempting to find various effective methods for improving their immunity to resist virus invasion and prevent virus infection by the new coronavirus. Under this background, some people take advantage of the panic mentality of the public to produce and disseminate a large amount of pseudohealth information on the Internet in the name of health. For instance, "drinking strong liquors can kill novel coronavirus," "drinking radix isatidis and smoking vinegar can prevent novel coronavirus," "drinking sterilizing fluid can kill novel coronavirus," and "wearing multilayer masks can prevent novel coronavirus." The publishers and disseminators of this pseudohealth information, with personal interests, act as "unhealthy" in the name of "healthy" and induce unwise behaviors of people who do not know the truth, which has brought great harm to the physical and mental health of the general public; additionally, it will cause property loss and life danger. All kinds of "Health Articles," "Cancer Alert," and "Private Sector" are filled in WeChat's circle of friends. Not only on social platforms, but also the whole network environment highlights a serious problem: health information is full of all kinds of health care pseudoscience, and the information is enough to make some people who lack health knowledge and literacy believe this kind of pseudohealth information. In addition, pseudohealth information spreads unscrupulously in rural areas, resulting in a series of serious consequences. For example, in recent years, there have been activities to promote fake health care products in rural areas

in China. The swindlers take advantage of the mentality of rural residents, such as seeking cheap prices and worrying about health, to carry out swindling actions which results in heavy losses to farmers. In 2015, Financial Channel of China Central Television reported that acetochlor pesticide residues were detected in strawberries, and long-term consumption would cause cancer risks. For this kind of pseudohealth information, it is difficult for nonprofessionals to distinguish whether the information is true or false. Although professionals interpreted that dosage determines toxicity with eight validation samples, it still caused a large scale of unsalable strawberries and brought great economic impact to farmers. Therefore, effective identification of pseudohealth information in networks is of great significance for maintaining the physical and mental health of the general public.

At present, there is no universally accepted definition of "pseudohealth information" in society. In general, pseudohealth information is interpreted as false health information without a factual basis, but in the real world, much pseudohealth information is fabricated based on certain facts, which only extends, distorts, exaggerates, and even fabricates the facts. Therefore, the pseudohealth information to be studied in this paper is that fabricated without a factual basis or with a certain factual basis but distorted or exaggerated by the publisher, the so-called health information that deviates from the truth. Network pseudohealth information refers to false health information that is fabricated or distorts the truth transmitted specifically through social media on the network. It is the "noise" in health transmission; it often induces people to form incorrect health cognition and even engage in improper health behaviors, which brings inestimable harm to the public's physical and mental health. Thus, it is of great practical significance to study the identification methods of network pseudohealth information to prevent the spread of pseudohealth information and maintain social stability.

2. Related Works

Internet pseudohealth information mostly belongs to the nature of rumors, which have the characteristics of rapid spread, wide influence range, and great social harm. It often induces a wide range of network public opinions or public health events and attracts widespread attention. At present, the research on pseudohealth information identification mainly focuses on the following three aspects: (1) The "select instance" (or sliding window) classification method. For example, Molinaro and Greco proposed a two-stage instance selection algorithm, which is divided into two stages: concept detection and retraining. If the semantics of class health are detected, the algorithm will automatically update the classifier and find classification labels in class health information data for classification [1]. Han et al. proposed the sliding window algorithm, which can deal with the attribute classification problem of network pseudohealth information [2]. Hoens et al. proposed an support vector machine (SVM) model for detecting network pseudohealth information, and the classification of network pseudohealth information was realized by updating the weight allocation of instances [3]. (2) The batch classification method. For example, Sutskever et al. proposed the information batch processing model, which realizes batch processing of class health information by constantly updating the classifier, thus realizing the classification of pseudohealth information [4]. Rodriguez and Laio proposed an integrated model based on time limits, which can preliminarily compare and distinguish pseudohealth information and health information in the network [5]. (3) The classification method of online learning. For example, the pseudohealth information network online learning combination model proposed by Brzezinski and Stefanowski is composed of network online classifiers. Since the number of classifiers is usually fixed, as a result, the weighted sum update is also fixed [6]. Shi et al. proposed an online incremental algorithm to deal with the classification of network pseudohealth information. Due to the narrow value of online increments, which leads to poor fault tolerance [7], Eskandari and Javidi adopted the network online learning method to classify pseudohealth information through centralized processing, but its classification accuracy was relatively low, and the classification effect was also poor [8].

In previous related studies, scholars have proposed a variety of classification algorithms for the identification of network pseudohealth information, including the combination model of different algorithms. These algorithms and models have good recognition effects on pseudohealth information with obvious identification of information sources and text semantic tags. However, it is difficult to identify pseudohealth information with unclear information sources and unclear semantic tags in the network and is also difficult to detect and classify. In the previous research on pseudohealth information identification, whether it is "select instance" (or sliding window) classification method, batch classification method, or the online learning classification method, each has its own advantages and disadvantages. Although pseudohealth information can be classified from different aspects, the existing methods are mainly single classifiers or batch processing, which result in either the classification cannot be effective or the recognition accuracy not being high. Through the research on pseudohealth information, this paper aims to help people distinguish pseudohealth information and improve their health information literacy, thus fundamentally improving the quality of network health information and purifying the network health information environment. Based on this, this paper proposes an integrated combination of the latent Dirichlet allocation (LDA) algorithm and data partitioning and accurate update. By identifying network pseudohealth information by topic, class tag blocks are accurately updated and integrated with data blocks to effectively identify and classify pseudohealth information.

3. Research Methodology

3.1. Concept Definition. The combination algorithm proposed in this paper to identify the problem of network pseudohealth information, the core of which is to divide the

dataset corresponding to network pseudohealth information into "granularity" blocks according to its class label properties. To detect the minimum information unit attribute contained in the dataset, the dataset is continuously updated in blocks according to the category of information attribute contained in the minimum information unit and is reintegrated and classified according to the category of data blocks, to effectively identify pseudohealth information. The concepts involved in this combined algorithm are as follows.

Definition 1. Pseudohealth information (semantic definition). The so-called pseudohealth information refers to misleading others to follow blindly or accept false publicity in a misleading and deceptive way in the name of health to realize the personal interests of producers and broadcasters and has been falsified.

In summary, pseudohealth information usually appears in the external form of health information. It takes advantage of people's demand for health information and uses false, deceptive, misleading, and other ways and means to spread and advocate unscientific, false content to achieve personal purposes, and the information has been falsified. The semantics of pseudohealth information deviate from the information title and semantic label and have a conceptual drift with the original meaning. According to this, pseudohealth information can be defined in terms of information from the perspective of information dissemination, and its information definition is shown in Definition 2.

Definition 2. Pseudohealth information (information definition). Class health information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$, where x is the attribute value and y is the vector of the class label, decomposes its joint probability $P(x, y)$ into $P(x, y) = P(x)P(y|x)$. If the prior probability $P(x)$ and conditional probability $P(y|x)$ of the sample in the class health information dataset change, semantic concept drift occurs in the class health information dataset S: during semantic concept drift, if $P(x)$ does not change and $P(y|x)$ changes, it belongs to the concept drift of the conditional change class; that is, the class health information is determined as true health information; if $P(x)$ and $P(y|x)$ change, it belongs to the concept drift of feature change; that is, similar health information is false health information; that is, it is determined as false health information.

Generally, health information refers to similar health information in which the attributes or labels contained in the information dataset have not changed, but their external representations or conditions have changed over a period of time; however, pseudohealth information refers to those that appear as "health" and have a relatively stable feature distribution. However, the class health information is changed or deviates from the class label corresponding to the "health" eigenvector.

Definition 3. Data block. If the information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$ is divided into sequences arranged in sequence $z_1, z_2, \ldots, z_i, \cdots, z_n, \ldots$, each sequence contains a data record or several logical markers; if each sequence $z_i = (x, y)$ consists of eigenvectors $x \in X$ and class labels $y \in Y$, sequence elements z_i are called data blocks.

Definition 4. Data block integration. If the information dataset $S = \{(x^t, y^t) | t = 1, 2, \ldots, T\}$ is divided into data blocks $z_1, z_2, \ldots, z_i, \cdots, z_n, \ldots$ with uniform size, each type of information data block contains d data blocks; for each newly added block z_j, the weight of the classifier $C_i \in \varepsilon$ is weighted by the weighting function $Q(\cdot)$. The weighting function $Q(\cdot)$ depends on the classification accuracy of the classifier. If the size of the data block set is set to k and does not exceed the limit, z_j is classified and added to a data block set of a certain type; if a data block set is a full set and the weight of the newly added data block is greater than that of the remaining data blocks, the newly added data block replaces the weakest block in the original set, and this process is called data block integration.

3.2. Algorithm Design

3.2.1. Algorithm Idea. The combination algorithm proposed in this paper blocks the data of the health-like dataset; that is, based on the class labels in the health-like dataset, topic recognition, information dataset partitioning, data block classification integration, and semantic offset detection are involved in the LDA model, Algorithm 1, SVM model, and Algorithm 2. The logical framework of the combined model is shown in Figure 1.

3.2.2. LDA Model. LDA was proposed by David Blei, Andrew Ng, and Michael I. Jordan in 2003. It is mainly used for document-topic generation and contains three levels of structure: document, topic, and word. Therefore, it is also called the probability model of the three-layer shellfish leaf stage [9]. As soon as the LDA model was proposed, it attracted the attention of scholars, especially in the field of semantic mining, which can greatly reduce the representation dimension of the text, thus making the model widely used [10, 11]. Additionally, as a typical representative unsupervised model, the LDA model has the advantage that the number of topics can be determined as long as important input parameters in the model are determined; therefore, the algorithm process is greatly simplified [12]. Based on this, when determining the optimal value of the number of document topics, this paper selects perplexity as an index to evaluate the pros and cons of the model, and its calculation equation is as follows:

$$\text{Perplexity}(D) = \exp\left[\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right]. \quad (1)$$

In the equation, M is the number of documents, D is the set of words in the document, w_d is the word, N_d is the number of words, and $p(w_d)$ is the probability of words in the document.

According to the statistical results, users who have published more Weibo information basically do not have the behavior of spreading fake health information, and their user credibility can be measured by the number of fans, the

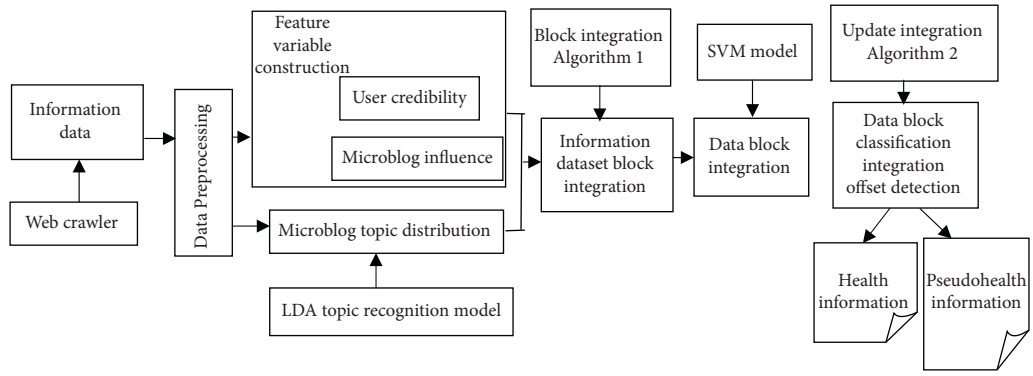

Figure 1: Logical framework of combined models.

number of followers, and the ratio; for those users who observe more but have fewer fans, the credibility is relatively low, and their fans are often the Internet Water Army. These users are most likely to be publishers or sources of a large amount of pseudohealth information. They publish or disseminate pseudohealth information through various network social platforms, such as Weibo and WeChat. Therefore, user credibility Reliability (u) can be defined by

$$\text{Reliability}(u) = \ln(\text{follower} - \text{following} + \text{num}) + \text{verify}. \tag{2}$$

In the equation, follower, following, and num are the number of fans, the number of followers, and the number of Weibo posts, respectively, after z − score standardization.Reliability (u) is an important basis for measuring the user credibility. The larger the value of Reliability (u) is, the higher the user credibility is.

For users, the number of fans, the number of Weibo forwards, the number of comments, and the number of praises are the basis for evaluating their influence. Generally, the more fans a user has, the greater the probability that the microblog posted by the user will be seen and spread by others and the more the corresponding forwarding, comments, and praise are. Regardless of what kind of Weibo operation behaviors fans perform, they all focus on the content published by users. Therefore, the influence of users' Weibo Influence (t) can be defined according to the following equation:

$$\text{Influence}(t) = \ln(\text{follower} + \text{repost} + \text{comment}) + \text{like}. \tag{3}$$

In the equation, follower, repost, comment, and like are the number of fans, forwarding, comments, and praise after z − score standardization. Influence (t) is an important indicator for evaluating the influence of users' microblogs. The greater the value of Influence (t) is, the greater the influence of user microblogs is.

3.2.3. Data Block Update Integration Algorithm

(1) Dataset Partitioning Algorithm. The identification of network pseudohealth information determines the essence of information semantics according to the deviation degree between the target class label and semantic ontology. If the semantic concept in information dataset S^t is replaced by S^{t+1} and the type of deviation is a subversive deviation, the "health" information content contained in the information semantics is replaced by pseudohealth information, its information semantic ontology has undergone fundamental changes, and the semantic ontology of network pseudohealth information belongs to this category. According to this principle, the information dataset S is now divided into data block streams $z_1, z_2, \ldots, z_i, \ldots, z_n, \ldots$, and each data block contains one record or several logical records. Classifier C_i is constructed, and the newly added data block Z_j is empowered. The classification performance of classifier C_i is determined by the weighted function $Q(\cdot)$. In the process of information dataset partitioning, if a certain type of data block set is not a full set, data block Z_j is added to this type of set; if a set of data blocks is a full set and the weight of block Z_j is greater than that of any block, the weakest block is replaced. The block integration algorithm of the dataset is shown in Algorithm 1.

(2) Data Block Set Classification Integration SVM Model. SVM is a typical representative binary classification model that is superior in classification generalization ability; therefore, it has been widely used in the field of information and data classification [13, 14]. In this paper, when identifying network pseudohealth information, the SVM model is adopted to integrate and classify the data block set to transform the instance sample dataset into the problem of solving convex quadratic programming. Then, the best classification hyperplane of the sample space is obtained. The classification hyperplane equation is as follows:

$$\omega^T \cdot x + b = 0. \tag{4}$$

In the equation, $\omega = (\omega_1, \omega_2, \ldots, \omega_7)$ is the normal vector, which determines the direction of the hyperplane; b is the displacement item, which determines the distance between the hyperplane and the origin; and $x = (f_1, f_2, \ldots, f_7)$ is the eigenvector of the sample point. The distance of the hyperplane is a controllable factor that makes the distance between two types of sample points and the classification hyperplane reach the optimal size based on the requirement of classification accuracy [15]. In addition,

the SVM model has good fault tolerance in the training process, and the optimal solution form of its optimal classification hyperplane equation is as follows:

$$\min_{\omega,b} \frac{\|\omega\|^2}{2} + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \begin{cases} T_i\left(\omega^T \cdot x_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ i = 1, 2, \ldots, N. \end{cases} \tag{5}$$

In the equation, x_i is the eigenvector of the i-th sample point, ξ_i is the relaxation variable of the i-th sample point, T_i is the category label of the i-th sample point, N is the number of training samples, and C is the penalty coefficient. The classification performance of the SVM model is determined by its kernel function. Choosing different kernel functions will lead to great differences in classification accuracy. At present, the kernel functions commonly used in SVM models include linearity, polynomial, and radial basis function (RBF) [16]. Since the classification accuracy of the RBF kernel function is much higher than that of other kernel functions and is suitable for situations where the number of features is less than or equal to the number of samples [17, 18], this paper chooses the RBF kernel function, as shown in the following equation:

$$K(x, z) = \exp\left(-\gamma\|x - z\|^2\right). \tag{6}$$

When the SVM model is classified and trained, the penalty coefficients C and γ in the RBF kernel function need to be determined in advance, the fault tolerance of the model is controlled by the former, and there is a negative correlation between the two; that is, the larger the penalty coefficient C is, the smaller the fault tolerance is. When C is too high, the overfitting phenomenon occurs [19, 20]; however, when C is small, to a certain extent, the classification accuracy of the model will be reduced accordingly. In other words, the parameter γ in the RBF kernel function affects the distribution of sample points mapped to high-dimensional space and exerts an influence on the penalty coefficient C, thus making the SVM model have high classification integration accuracy.

(3) Semantic Offset Detection Algorithm. The semantic changes in network pseudohealth information are very complex. Existing studies use online weighted and incremental classification methods to detect the changes in target semantics in network pseudohealth information, but data block integration is much more complicated than incremental classification, and the existing semantic offset detection algorithms have defects. To compensate for this defect, this paper adopts the semantic offset detection algorithm. The principle of this algorithm is that each data block contains one record or multiple logical markers, the data block set classified by the SVM model needs to be batch processed, and the candidate classifier corresponding to the integration component of the data block set is triggered for a classification check. If the current data block set is correctly classified, the original data block set of classification integration can be kept unchanged; if the fault tolerance of the current data block classification integration is poor or the classification accuracy is low, the integration component is reweighted, and the class tags in the data block are redetected to improve the classification accuracy of the classifier to effectively detect the target semantic attributes. Therefore, the semantic offset detection algorithm is shown in Algorithm 2.

4. Experimental Results and Analysis

4.1. Instance Data Acquisition. In this paper, crawler software is used to crawl experimental data, and pseudohealth information published by the publicity section of the Sina Weibo community management center is used as a reference. This pseudohealth information is reported due to false information and has been clearly confirmed by the government as pseudohealth information. Due to the spread of various pseudohealth information during the new coronavirus epidemic, the pseudohealth information in Sina Weibo is considerable. This paper crawled pseudohealth information from the API of Sina Weibo from February 1 to March 31, 2020, and randomly collected 1,183 pseudohealth information points. Among them, 759 of the original Weibo have more than 100 comments. The content of each Weibo was marked, counted, and sorted by its number of forwards, comments, and praises, and the experimental case dataset was constructed together with user information and the number of followers and fans.

To prevent the classifier from dividing all experimental data into health information, we added a manual verification step and selected some microblogs with comments greater than 100 and text, not pure symbols, and length greater than 10. The classification basis was obtained by means of manual verification technology and compared with health information. A total of 368 pieces of health information data were obtained through layer-by-layer screening, with more than 96.43 million comment texts. Based on the characteristics of comment anomaly parameters and SVM model parameters determined by the algorithm, this paper manually labeled the collected instance datasets. The selected instance dataset includes 359 pieces of pseudohealth information and 268 pieces of health information. When verifying the pseudohealth information recognition model, we made full use of the remaining 100 pieces of pseudohealth information and 100 pieces of health information to conduct precision comparison training experiments. The dataset composition of the experimental examples is shown in Table 1.

4.2. LDA Topic Recognition and Preprocessing. According to the data variables given in Table 2, the LDA model was used to preprocess the instance dataset to mine the document-topic distribution characteristics of the pseudohealth information dataset; the variables listed in Table 2 are the characteristic indicators required for LDA model preprocessing, and the meaning of each variable corresponds to

Input: S: an instance information dataset divided into blocks d; K: the number of members of the dataset; B: an instance buffer with a size of d; $Q(\cdot)$: classification quality measurement.
Output: ε: the integration of the classifier weighted as k.
(1) Information data block do for all $Z_j \in S$
(2) According to Z_j and $Q(\cdot)$, candidate classifier C' is established and empowered;
(3) According to Z_j and $Q(\cdot)$, all classifiers C_i in set ε are empowered;
(4) if $|\varepsilon| < k$, then $\varepsilon \longleftarrow \varepsilon \cup \{C'\}$;
(5) Else if $\exists i$: $Q(C') > Q(C_i)$, then replace the blocks in the weakest set with C';
(6) Initialize C' with B;
(7) $B \longleftarrow \varnothing$;
(8) Calculate the error of all types $d \in \varepsilon$ to S;
(9) Run the command on all instances of $Z_j \in \varepsilon$;
(10) End if
(11) End for

Algorithm 1: Dataset block integration algorithm.

Input: S: instance information data flow, D: information semantic offset detector, k: number of integrated members, B: instance buffer with size d, $Q(\cdot)$: classification quality measurement, t: number of instances;
Output: ε: offset detector integration with 1 classifier and k-class weighted classification;
(1) For all $x^t \in S$ instance do
(2) Gradually replace D with x^t
(3) $B \longleftarrow B \cup \{x^t\}$
(4) if $|B| = d$ or the offset is detected, then;
(5) According to W and $Q(\cdot)$, candidate classifier C' is constructed and empowered;
(6) According to $W$and $Q(\cdot)$, the classifier $C_i$in integration ε is empowered;
(7) if $|\varepsilon| < k$, then $\varepsilon \longleftarrow \varepsilon \cup \{C'\}$;
(8) Else if $\exists i$: $Q(C') > Q(C_i)$, then replace the weakest block in the integration with C';
(9) Initialize D;
(10) $B \longleftarrow \phi$;
(11) End if
(12) End for

Algorithm 2: Semantic deviation detection algorithm.

Table 1: Composition of experimental microblog datasets.

Category	Keyword	Number
Pseudohealth information data from February to March in 2020	Resistance viruses	217
	Immunity viruses	123
	Infection viruses	119
Health information data from February to March in 2020	Viruses	368

Table 2: Data variables.

Features	Document-topic distribution	User characteristics			Weibo features			
Feature metrics	0 ... n	Authentication	Number of fans	Number of attentions	Posted Weibo	Number of forwards	Number of comments	Number of praises
Variables	p_{mo} ... P_{mn}	Verify	Follower	Following	Num	Repost	Comment	Like

its characteristic indicators, where verif y_i indicates whether user u_i's Weibo account has been authenticated for personal information. If it was authenticated, u_i is 1; otherwise, it is 0. The characteristic indicators of other variables were consistent with the variable characteristic indicators in the user credibility and perplexity equation.

The result of LDA model preprocessing is shown in Figure 2. In the figure, the horizontal axis is the number of

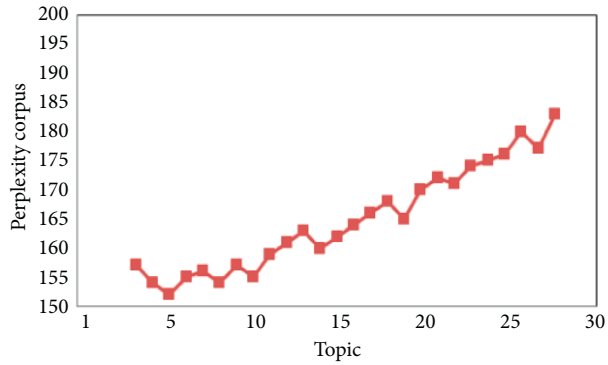

Figure 2: Perplexity-topic line chart.

Table 3: Distribution of subject words.

Topic 1	Alcohol	High temperature	Degree	Killing	Virus
Probability	0.043	0.039	0.037	0.015	0.009
Topic 2	Sterilizing fluid	Drinking	Virus	Killing	Effect
Probability	0.022	0.017	0.013	0.009	0.006
Topic 3	Mask	Multistory	Stopping	Stopping	Valid
Probability	0.075	0.046	0.033	0.028	0.025
Topic 4	Double *Coptis chinensis*	Inhibition	Virus	Mitigation	Treatment
Probability	0.049	0.039	0.027	0.021	0.021
Topic 5	5G	Spreading	Radiation	Carrying	Virus
Probability	0.036	0.023	0.023	0.008	0.005

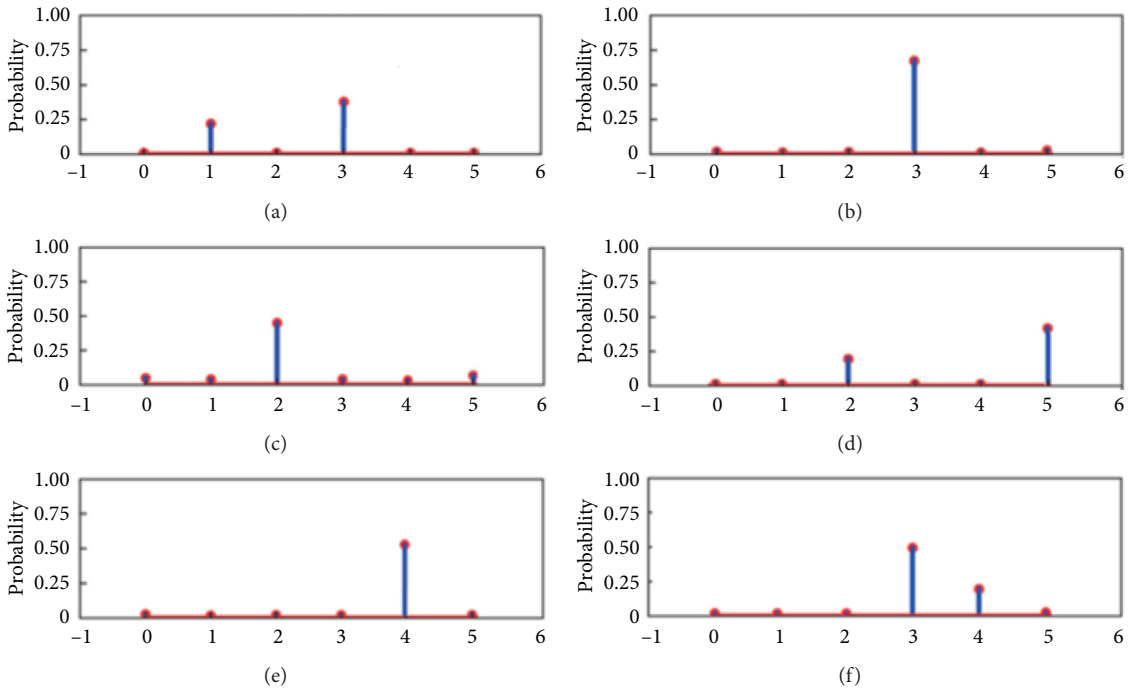

Figure 3: Document-topic distribution. (a) Topic, document 50. (b) Topic, document 61. (c) Topic, document 24. (d) Topic, document 39. (e) Topic, document 46. (f) Topic, document 45.

topics, the vertical axis is the perplexity, the polyline is 3 to 28, and the interval is 1. As seen in Figure 2, with the increasing number of subjects, the perplexity also continues to increase, but the rising track has certain volatility. When the number of subjects is 5, the perplexity reaches its lowest value. As the number of topics increases, the perplexity also

increases in a wave and reaches the maximum when the number of topics is 28. Based on the minimum principle of "perplexity + number of topics," 5 is selected as the topic parameter value of the LDA model.

After determining the optimal topic parameter value of the LDA model, the LDA model can be used to perform deep semantic training on the segmented instance dataset and then determine the distribution rules of "document-topic" and "topic-word" to determine the class labels or classification features of topics and words and prepare for the block and reintegration of instance datasets. The training results are shown in Table 3. As seen in Table 3, the results of LDA model training have obtained 5 topics. Now, the first 5 words are selected to represent each topic, and the probability of occurrence of each word is given.

Next, we randomly selected 6 documents as examples and show their "document-topic" distribution map to explore the probability of their themes and subject words. The specific results are shown in Figure 3. As seen in Figure 3, the probability of six document topics is different, but there is always a higher probability of one or two topics, while the probability of other topics is lower, which shows that the LDA model can divide the topic of microblog text well and provide a good foundation for the next step of this paper to block and integrate the microblog pseudohealth information instance dataset.

4.3. Integration of Dataset Partitioning and Classification

4.3.1. Block Experimental Datasets. The experimental dataset processed by the LDA model was cross-verified K times, and the instance dataset (S) was input. It was randomly divided into K subsets S' ($S' = \{S_1, S_2, \ldots, S_K\}$) with different sizes and mutual exclusion. In addition, S' was trained and tested K times; that is, in i iterations, subset S_i was retained as a test set, and the remaining subsets were used for training. The block efficiency is K iterations of training times divided by the total number of experiments. The$K$-fold cross-validation uses the classifier in Algorithm 1 to extract the weight of interactive information. The purpose of the cross-validation experiment is to verify the block efficiency and performance of Algorithm 1.

According to Algorithm 1, for a given instance dataset, if the attributes and class labels of the information text are obvious, the accuracy of the instance dataset is very high; if the attributes and class labels of the information text are vague or not clearly defined, the window algorithm (Algorithm 2) needs to be used to detect semantic deviation. In the process of grouping instance datasets, with the change in candidate classifier C', the classification discrimination boundary also changes. For all classifier C_i weights, the instance information dataset S is divided into data blocks of uneven size: $z_1, z_2, \ldots, z_i, \ldots, z_n, \ldots$; the candidate classifier C' is established according to Z_j and $Q(\cdot)$, and it is empowered accordingly so that the decision boundary of the instance dataset will not fall into the center point of one-dimensional, two-dimensional, and three-dimensional spherical Gaussian step by step, the cross-validation data blocks present Gaussian distribution, and the block discrimination boundary is composed of two hyperbolic surfaces. The block decision area is not simply connected but the area where the two elliptical contour lines formed by probability density are located, as shown in Figures 4(a) and 4(b).

In Figure 4(a), candidate classifier C' implements the partitioning of instance datasets according to the attributes of class labels. It uses all candidate classifiers C' in set ε to assign and update data blocks and creates k components to retain the original class labels of data blocks. The weight is updated based on the size of the instance buffer d to ensure that all data blocks have corresponding nonzero weights. In Figure 4(b), instance buffer d can not only retain the class tags of data blocks but also decide whether to replace the data blocks with the weakest class tags in the data block set according to classifier C_i. In addition, the data blocks with the weakest class labels can be removed or collected into the sets of other classes to effectively block instance datasets.

4.3.2. Data Block Classification and Integration. Vectorization is required for data block classification and integration. This paper uses the SVM model for classification and integration training, calls the libSVM tool, and adjusts the values of parameters C and γ to make the covariance matrix of the instance data block set distribution equal to obtain two n-dimensional spherical distribution information sets, namely, "health" and "pseudohealth" information data block classification integration datasets σ_1 and σ_2, where σ_1 and σ_2 are located on both sides of a $n-1$-dimensional normalized hyperplane. The hyperplane is the classification decision boundary of the two. The central line of the two n-dimensional spherical distributions formed by σ_1 and σ_2 is perpendicular to the hyperplane, as shown in Figure 5(a) and 5(b). In the process of classification integration, assuming $\exists i$: $Q(C') > Q(C_i)$, all classifiers C_i in the set ε are empowered according to Z_j and $Q(\cdot)$. If the errors of all types $d \in \varepsilon$ to S are equal, then all data blocks with $Z_j \in \varepsilon$ are classified and weighted. By measuring the Oushi distance from each data block to the ε-mean vector, the "minimum distance" of the boundary (hyperplane) is judged based on ε display and classification, classify and collect the weighted data blocks into the nearest dataset $\sigma_i (i = 1, 2)$, and the weakest data block in the set $\sigma_i (i = 1, 2)$ is replaced with C' to realize the preliminary classification and integration of data blocks, as shown in Figure 5(a).

The window algorithm (Algorithm 2) is different from other integrated classifiers. Its combination with the SVM model can continuously update and empower data blocks; therefore, data blocks are classified and integrated into the form of class labels, and the semantic deviation of data blocks can be effectively detected. The candidate classifier C' in Algorithm 2 and all the classifiers C_i in set ε determine the distance between the classification integration dataset and the superplane, which continuously updates the weight of the instance data block set d. The distance between the two types

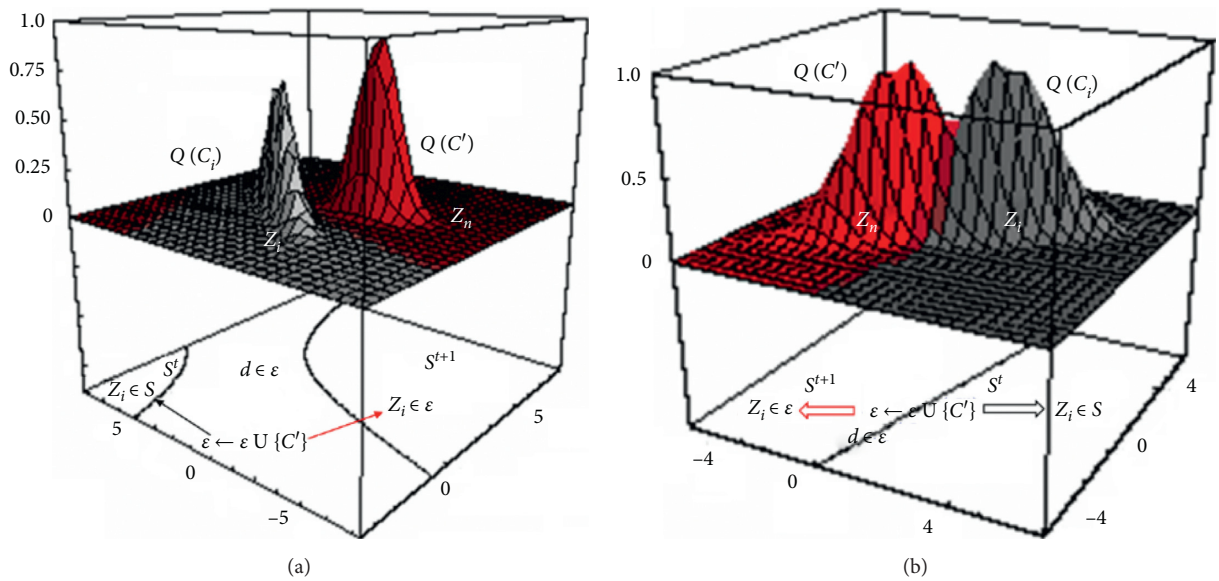

FIGURE 4: The process of instance dataset partitioning: (a) preliminary partitioning of candidate classifier C'; (b) weighting and updating of classifier C_i.

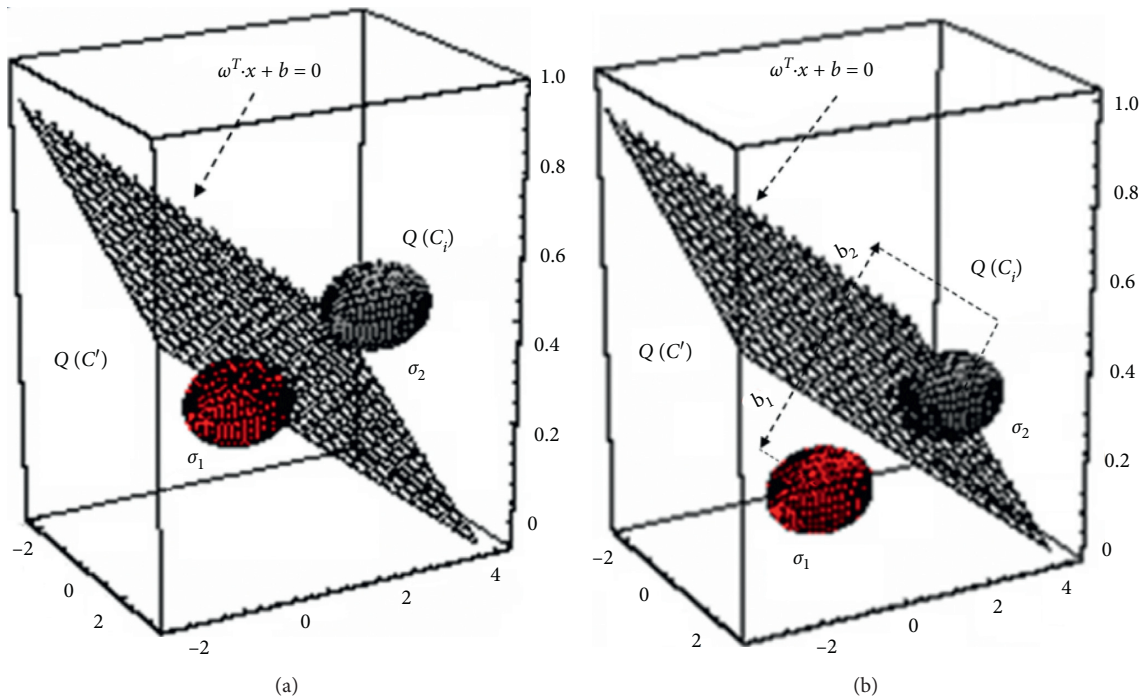

FIGURE 5: Classification integration process: (a) preliminary classification integration of data blocks; (b) precise classification integration of data blocks.

of datasets and the classification hyperplane is separated to the maximum through the best classification hyperplane in the SVM model (see equation (4)). At the same time, the slack variable ξ is introduced to improve the fault tolerance performance in the training process of SVM, and the sample points affected by the parameter γ in the RBF kernel function are mapped to the low-dimensional space to continuously correct the classification and integration efficiency so that the instance dataset can be accurately divided into "health" and "pseudohealth" information sets σ_1 and σ_2. The precise classification and integration process is shown in Figure 5(b).

4.4. Performance Evaluation of Classified Detection. To illustrate the advantages of the algorithm proposed in this paper, the logistic algorithm [21], decision tree (DT) [22],

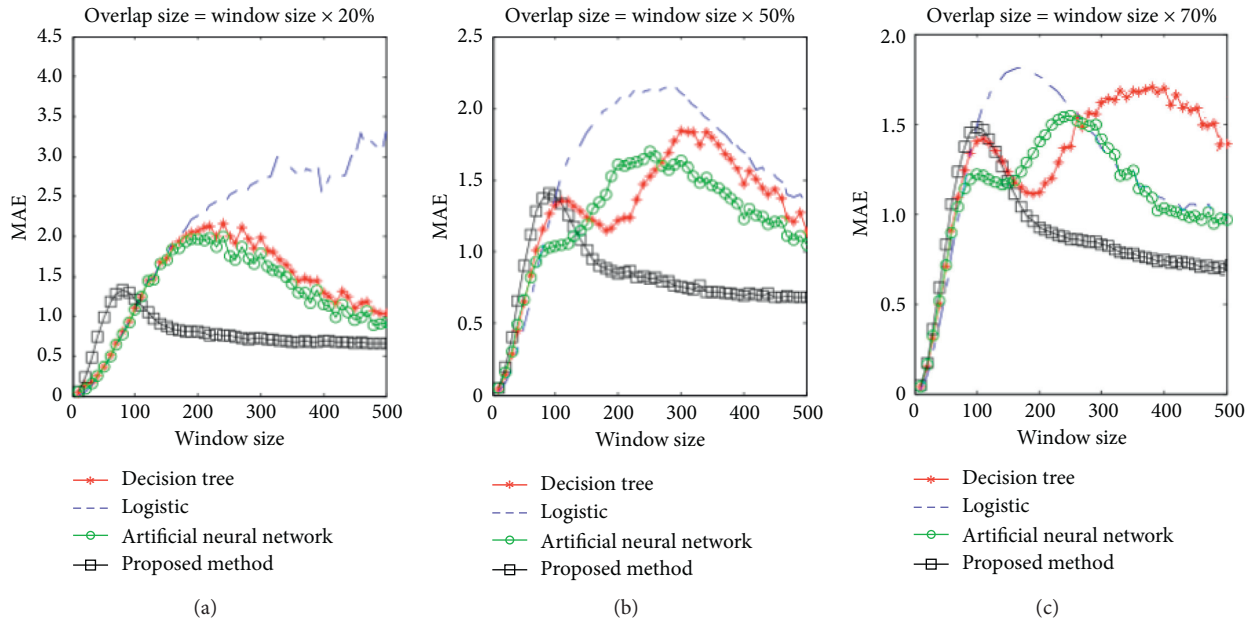

Figure 6: Detection performance of each classifier under different overlapping window size units: seconds.

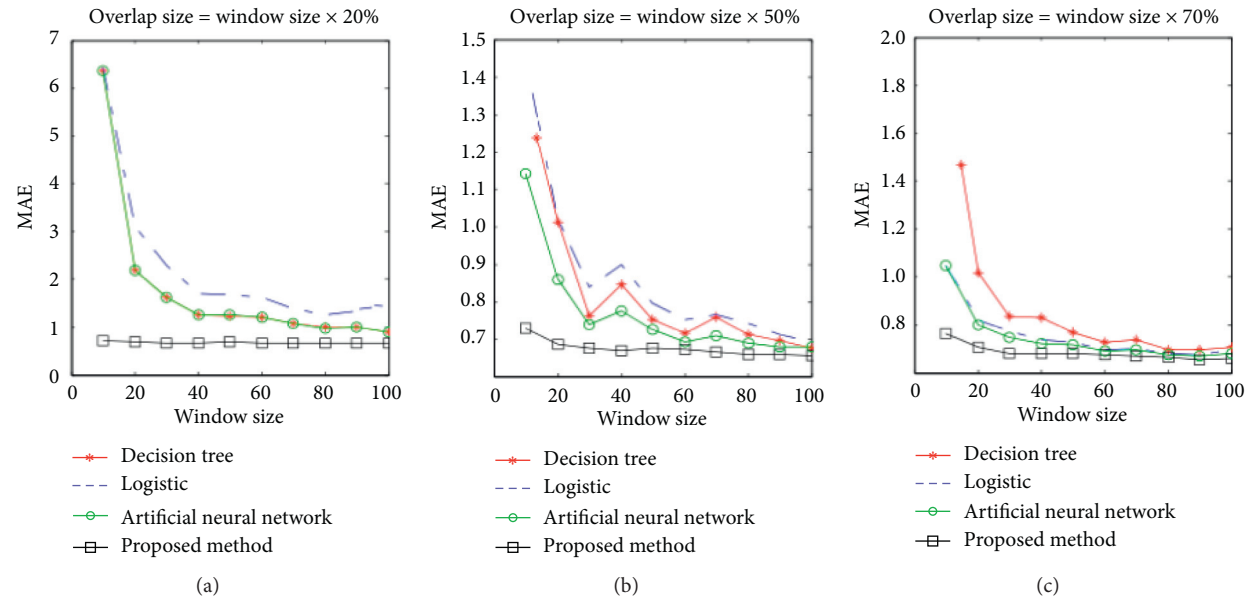

Figure 7: Detection performance of each classifier under different overlapping window sizes units: minutes.

and artificial neural network (ANN) [23] are now adopted for comparison. In addition, the classification accuracy among the four algorithms is tested. The classifiers of these four algorithms can update and classify the instance data block sets by using sliding windows in a free combination way. Therefore, the experimental data block set can be classified and integrated. Because the cross-validation strategy can overcome the overfitting of the classifier and enhance the generalization ability of the four algorithms, the classification accuracy of the four algorithms is compared by using the cross-validation strategy. The 10 instance subsets in this paper are randomly used for training to verify the classification accuracy of different models. The experimental results are shown in Figure 6.

As seen in Figure 6, the detection effects of the DT, logistic algorithm, and ANN are better than that of the method proposed in this paper within 0–100 seconds. However, the method in this paper is better than the other three algorithms in more than 100 seconds because the average absolute error (MAE) of the classification of the

Table 4: Classification accuracy of four classifiers units: %.

Data type	Classifier			
	DT	Logistic	ANN	Proposed method
Training sample	71.35	84.62	87.40	96.88
Sample test	76.29	86.07	91.18	98.73

method proposed in the paper is higher than that of the other three methods in 100 seconds; however, in more than 100 seconds, it is lower than that of the other three methods, and the three overlapping windows all have similar situations. To further illustrate this problem, the window unit is set to minutes, the sliding window size is 100 minutes, and the overlapping size is equal to 20%, 50%, and 70% of the window size. Four algorithms are used to detect the dataset of the example, and the detection effect is shown in Figure 7.

In Figure 7, the DT, logistic, and ANN algorithms can greatly reduce MAE by adjusting parameter settings and adopting supervised/semisupervised methods to increase the classification effect after 30 minutes. The algorithm in this paper can efficiently detect and classify instance datasets from the beginning, and its MAE value always fluctuates between 0.5 and 0.8. Therefore, whether in seconds or minutes, the algorithm in the paper is obviously superior to the DT, logistic algorithm, and ANN model.

The classification accuracy of the four algorithms is compared with the example dataset in the paper. The experimental results are shown in Table 4. As seen in Table 4, the classification accuracy of the four classifiers is quite different: the classification accuracy of the algorithm in the paper is the highest, the classification accuracy of the training sample is as high as 96.88%, the classification accuracy of the test sample is 98.73%, DT has the lowest classification accuracy, the classification accuracy of its training sample and test sample is 71.35% and 76.29%, respectively, the accuracy of the logistic algorithm and ANN is between the two, and the precision of ANN is slightly higher than that of the logistic algorithm.

5. Conclusions

The identification of network pseudohealth information is not only the frontier and focus in the field of news dissemination but also the focus and difficulty in the field of data mining. Although some scholars have studied this problem and proposed many recognition methods, the existing methods are mainly single classifiers or batch processing, which result in the fact that either the classification cannot be effective or the recognition accuracy is not being high. Based on the class tag attributes of network pseudohealth information datasets, the paper proposes a combination algorithm integrating data partitioning and classification update based on previous research results, integrates LDA topic recognition model, dataset partitioning algorithm, SVM data block classification integration model, semantic offset detection algorithm, and other methods, and adopts Web crawler technology to conduct simulation experiments based on the pseudohealth information of the Sina Weibo platform during the epidemic from February 1 to March 31, 2020. The simulation results show that the combination algorithm proposed in this paper has good superiority in both the subject recognition of pseudohealth information and the block and integration classification of instance datasets. Compared with DT, logistic algorithm, and ANN, the experimental results show that the classification integration accuracy of this method is higher than that of these three methods, which fully illustrates the reliability and practicability of the method in the paper. The identification of pseudohealth information in the future is of great significance for maintaining normal public health order and building a "Healthy China." Traditional mainstream media has high authority and influence. As a public tool of the society, media should perform its functions to serve the audience and the society and strengthen the check of fake health information to clarify its authenticity. At the same time, the media should also clarify the pseudohealth information that disturbs people in time to prevent the spread of pseudohealth information, which is also a way for the media to maintain their own image and authority. Therefore, we should not only pay attention to the problems existing in the dissemination of various kinds of information, but also make full use of technical means and tools to curb the further dissemination and influence of pseudohealth information.

Data Availability

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

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the general program of the Natural Science Foundation of Shandong Province (no. ZR2019MG021) and the Key Projects of the National Statistical Scientific Research Plan (no. 2019LZ19). The research was also supported by the social science planning (dominant discipline) research project of Shandong Province (no. 19BYSJ19).

References

[1] C. Molinaro and S. Greco, "Polynomial time queries over inconsistent databases with functional dependencies and foreign keys," *Data & Knowledge Engineering*, vol. 69, no. 7, pp. 709–722, 2010.

[2] J. W. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*pp. 327–330, Morgan Kaufmann, Burlington, MA, USA, Third edition, 2011.

[3] T. R. Hoens, R. Polikar, and N. V. Chawla, "Learning from streaming data with concept drift and imbalance: an overview," *Progress in Artificial Intelligence*, vol. 1, no. 1, pp. 89–101, 2012.

[4] I. Sutskever, J. Martens, U. Dahl, and U. E. Hinton, "On the importance proceedings of initialization and momentum in deep learning," in *Proceedings of the International Conference on Machine Learning*, pp. 1139–1147, Atlanta, GA, USA, June 2013.

[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6, pp. 1492–1496, 2014.

[6] D. Brzezinski and J. Stefanowski, "Combining block-based and online methods in learning ensembles from concept drifting data streams," *Information Sciences*, vol. 265, no. 5, pp. 50–67, 2014.

[7] Y. Shi, F.-L. Chung, and S. Wang, "An improved TA-SVM method without matrix inversion and its fast implementation for nonstationary datasets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 2005–2018, 2015.

[8] S. Eskandari and M. M. Javidi, "Online streaming feature selection using rough sets," *International Journal of Approximate Reasoning*, vol. 69, no. 2, pp. 35–57, 2016.

[9] T. Gocken and M. Yaktubay, "Comparison of different clustering algorithms via genetic algorithm for VRPTW," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 574–585, 2019.

[10] J. Qu, Z. Ji, C. Lin, and H. Yu, "Fast consensus seeking on networks with antagonistic interactions," *Complexity*, vol. 2018, Article ID 7831317, 15 pages, 2018.

[11] D. Kurunathan, S. Shanmugathas, and K. Ashoka, "Analysis of relation between customer behavior and information technology market," *Journal of System and Management Sciences*, vol. 9, no. 1, pp. 87–104, 2019.

[12] Z. Han, X. Liu, and J. Kou, "Interdisciplinary subject recognition based on Rao-Stirling index and LDA model-a case study of nanotechnology," *Information Science*, vol. 38, no. 2, pp. 116–124, 2020.

[13] Y. Zheng, X. Hu, and J. Yin, "Health data fusion method based on multi-task support vector machine," *System Engineering Theory and Practice*, vol. 39, no. 2, pp. 418–428, 2019.

[14] Y. Yang, F. Zhang, and H. Xue, "A modal Fu Liye-support vector machine optimization method for abnormal data reconstruction of water intake monitoring," *Operations Research and Mamagement Science*, vol. 28, no. 2, pp. 52–59, 2019.

[15] X. Li, S. Wu, X. Li, H. Yuan, and D. Zhao, "Particle swarm optimization support vector machine model for machinery fault diagnoses in high-voltage circuit breakers," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 6, pp. 1–10, 2020.

[16] K. Bi and T. Qiu, "An intelligent SVM modeling process for crude oil properties prediction based on a hybrid GA-PSO method," *Chinese Journal of Chemical Engineering*, vol. 27, no. 8, pp. 1888–1894, 2019.

[17] Y. Lu, W. Wei, Y. Li, Z. Wu, and H. Jin, "The formation and evolution of interorganisational business networks in megaprojects: a case study of Chinese skyscrapers," *Complexity*, vol. 2020, Article ID 2727419, 17 pages, 2020.

[18] R. Goyat, G. Kumar, M. K. Rai, and R. Saha, "Implications of blockchain technology in supply chain management," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 92–103, 2019.

[19] S. Natalija and S. Dragan, "Accelerating multiple flow accumulation algorithm using MPI on a cluster of computers," *Studies in Informatics and Control*, vol. 29, no. 3, pp. 307–316, 2020.

[20] T. Saric, G. Simunovic, D. Vukelic, K. Simunovic, and R. Lujic, "Estimation of CNC grinding process parameters using different neural networks," *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 6, pp. 1770–1775, 2018.

[21] M. Li and H. Xu, "Reliability window analysis of Gap zero gate based on Logistic model," *System Engineering Theory and Practice*, vol. 39, no. 2, pp. 531–538, 2019.

[22] T. Chen and L. Zhu, "Assessing the performance of Decision tree and neural network models in mapping soil properties," *Journal of Mountain Science*, vol. 16, no. 8, pp. 1883–1847, 2019.

[23] L. Macyszyn, C. Jedryczka, and R. Staniek, "Design and finite element analysis of novel two-stage magnetic precession gear," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 586–595, 2019.