Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4927504, 10 pages
https://doi.org/10.1155/2022/4927504

Research Article

Computer Network Confidential Information Security Based on Big Data Clustering Algorithm

Weigang Liu

College of Information Engineering, Tianjin Modern Vocational Technology College, Tianjin 300350, China

Correspondence should be addressed to Weigang Liu; liuweigang1981@stu.wzu.edu.cn

Received 15 March 2022; Revised 21 April 2022; Accepted 28 April 2022; Published 9 June 2022

Academic Editor: Jun Ye

Attacks on network systems are becoming more and more common, the current state of increasingly sophisticated attack methods, the emergence of intrusion prevention technology is the inevitable result of the development of computer technology and network technology, and research on intrusion prevention has become a new focus of network security technology research in recent years. In order to ensure the security of computer network confidential information, the authors propose a semisupervised clustering intrusion detection algorithm. An overview of machine learning, followed by an explanation of the theory of cluster analysis, simulation experiments were carried out using the K-means algorithm and the semisupervised clustering algorithm proposed by the author, for 10,000 records, the K-means clustering algorithm and the semisupervised clustering algorithm described in this paper are used, respectively, and intrusion detection data tests were performed. At the same time, different K values were selected, three datasets were selected from "kddcup.newtestdata_10_percent_corrected," the test data were tested separately, and their average value was taken as the test result. From the simulation results, the detection rate of the semisupervised clustering algorithm is higher than that of the K-means clustering algorithm, and the false alarm rate and K-means algorithms have also been improved. Therefore, the author's semisupervised algorithm enhances the stability of the system, and the performance of the K-means algorithm is improved to a certain extent. When the value of K gradually increases, the false alarm rate also increases; however, when K is 20, the detection rate is maximized, from this, it can be known that when K is 20, its detection rate reaches 91.76%, and the false alarm rate is 8.54%. The detection rate of the author's algorithm is significantly higher than the other two algorithms, the false positive rate is slightly higher than K-means, and the false positive rate is lower than that of the other algorithm, proving the superior performance of our algorithm.

1. Introduction

With the continuous updating of computer Internet technology, it has gradually been applied in various fields, it has had a huge impact on people's lives and work, a large amount of data information has exploded, and it involves a lot of private information. Aerospace transportation, as an extremely important information industry in the development of modern society, it is related to classified information security and aerospace security and is also closely related to the stable development of social security. The spread of the Internet, so that everyone will be exposed to the Internet in their daily lives, personal information data security also has certain threats and risks. Therefore, strengthening the management and protection of computer network information security is a very important project [1]. In the process of designing and implementing computer network information security protection, Bayesian classification algorithm is one of the most common methods and has achieved significant research results, considering the growing maturity of big data clustering algorithm, the scope of application is also getting wider and wider, and more well-known technical achievements are gradually born [2]. At present, the research on computer network information security protection strategies is still in its infancy, especially the correlation between different target attributes, the

proportion of nonlinear relationship occupies more than half, if the conventional method is used, it is difficult to fully reflect the actual relationship, and in the process of analysis, there will also be contradictions, and disorders are likely to occur. In the process of computer network information transmission, because the network itself has certain security risks, this is also a relatively common information security risk, it is also affected by human factors, there will be more and more unsafe factors in the existence of information [3]. The most common network information security problems are as follows: the first is the vulnerability of TCP/IP, the second is the insecurity of the network structure, the third is the possibility of information being stolen, and the fourth is the weak awareness of safety management of relevant staff [4]. Computer network technology is inseparable from the TCP/IP protocol, TCP/IP is vulnerable, the main reason is that the protocol does not pay enough attention to computer network security issues, and the TCP/IP protocol reflects the openness of the network more. This also gives attackers an opportunity to take advantage of, attackers open up the environment through the network, find loopholes, and attack the network, resulting in various information security problems. The computer network system has security risks, mainly because the network system is composed of countless local area networks, this also makes the computer network very large, if there is a communication behavior, there may be a risk of being attacked, the attacker only needs to pass a host, can operate, steal information and data, and continue to carry out the next attack. Computer network information has great security risks and is easy to be stolen, the main reason is that in the computer network, a large amount of data information has not been encrypted, when users use the computer network, they also use some free software, there will be some security risks, and gives attackers an opportunity to exploit vulnerabilities to eavesdrop. During the operation of computer users, due to the lack of security awareness, there will also be greater security risks, although the current computer network also has certain security protection mechanisms and measures; however, users do not make full use of security protection mechanisms and measures [5]. For example, some users think that the firewall is very troublesome and affect their use of some software, so they choose to close the firewall, in the case of not obtaining the authentication of the firewall proxy server, and the PPP connection, as a result, the firewall is useless, and the potential security risks may break out at any time [6]. There are many factors that threaten the security of computer network information, among which, hacking is one of the most common factors, hackers are one of the biggest threats to modern computer network systems, if the network is attacked by hackers and the server is damaged, it cannot provide normal services for users, as a result, the network is paralyzed, resulting in very serious consequences. Intrusion prevention system combines the functions of intrusion detection and firewall, it can monitor network traffic, timely interruption, adjustment or isolation of intrusion behavior is an active and resourceful intrusion defense system [7]. Figure 1 shows the application deployment of intrusion prevention.

2. Literature Review

In recent years, with the improvement of hardware performance and the gradual improvement of medical networks, a large number of breast cancer diagnosis and treatment plans based on machine learning algorithms have been proposed, for example, Xiong et al. applied SVM classifier and Naive Bayes classifier to diagnose breast cancer data [8]. Kim et al. propose a breast cancer diagnosis scheme that mixes k-means clustering algorithm and SVM classifier. The scheme first uses the k-means clustering algorithm to identify the hidden patterns of benign and malignant breast cancer and set up an effective value to measure the clustering effect to determine the optimal number of clusters. Then, the similarity between the data and the hidden pattern is measured by the membership function, the dimensionality reduction of the data is realized, and finally, the breast cancer data is classified and diagnosed through the SVM classifier [9]. Al-Salhi et al. based on logistic regression algorithm and decision tree algorithm, analysis of prognostic factors for breast cancer [10]. Qian et al. proposed a safe outsourcing drug detection system based on DT-PKC encryption system, the system classification is completed by the SVM algorithm, and privacy protection is achieved by designing a secure multiparty protocol [11]. Yao et al. based on the fully homomorphic encryption algorithm, in the ciphertext domain, the minor allele frequencies and chi-square detection in genome-wide association analysis are calculated and support the calculation of Hamming distance and approximate edit distance for encrypted DNA sequences [12]. Meng et al. based on the BGV fully homomorphic encryption scheme, a secure mobile medical model using edge computing technology is proposed, provide health data diagnosis and analysis for members of the smart medical system, the calculation types are k-means clustering and fuzzy c-means clustering [13]. Monteith et al. proposed differential privacy protection technology, which is a strictly defined privacy protection model, it is resistant to all attacks based on background knowledge assumptions. The main principle is to achieve privacy protection by adding random noise to distort the original data, but at the same time, it does not affect the statistical properties of the original data [14]. Kirubakaran and Ilangkumaran proposed a fully dynamic fully homomorphic encryption scheme, and the scheme no longer limits the number of allowable parties [15]. Based on the BGV encryption scheme, Ge et al., a BGV-type multikey fully homomorphic scheme CZW is constructed, and the security of the scheme is based on the LWE problem on the ring. Compared with the GSW-type multikey fully homomorphic scheme, the CZW scheme has simpler ciphertext expansion, and the encryption method is no longer by bit encryption, but can encrypt ring elements. In addition, the scheme also supports technical acceleration such as batch processing, and the computational efficiency is higher than that of the GSW-type multikey fully homomorphic encryption scheme [16]. Based on the BCP encryption algorithm with double decryption mechanism, Cui et al., an outsourced secure multiparty computation framework is proposed. The scheme uses two noncollusion cloud servers,

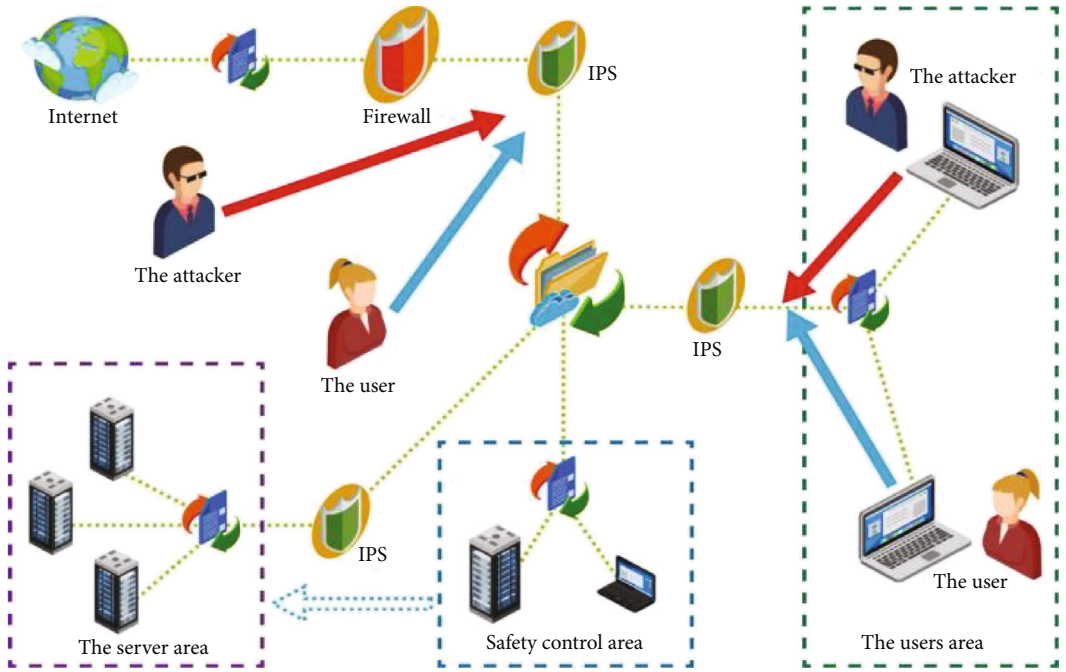

Figure 1: Application deployment of intrusion prevention.

one for storage and computing, a master private key that contains decryptable user data, realizes computing outsourcing by constructing a secure protocol [17]. This paper is an overview of machine learning, followed by an explanation of the theory of cluster analysis. Analyzing the author's intrusion detection method, and comparing it with several other common intrusion detection algorithms, for example, the correct can be established to improve the detection accuracy, the false alarm rate is reduced, and the robustness of the system is enhanced. In the author's algorithm, a small number of labeled samples provide correct guidance for the initial formation of normal and abnormal cloud models. The dynamic weighting method is used to solve the problem that high-level data is difficult to process, enable the data to learn from each other, and gradually form a relatively stable cloud model, and over-reliance on prior knowledge of the data is avoided. In contrast, the performance of the author's algorithm relative to the general clustering algorithm has greatly improved, for a certain extent, it solves some problems existing in the current intrusion detection.

3. Methods

3.1. Overview of Machine Learning. With the development of computers, machine learning has penetrated into many fields such as pattern recognition, data mining, and computer graphics. Machine learning is based on different classification basis, and there are different classification methods. According to whether the training sample data used in the training process has label information, machine learning can be divided into unsupervised learning, supervised learning, supervised learning, and semisupervised learning [18].

3.1.1. Unsupervised Learning. Unsupervised learning is a kind of unclassified data information for analysis and recognition, at the same time, cluster-related knowledge can be applied to unsupervised learning, in order to analyze the sample data and predict the category information of the sample. In unsupervised learning, a set of known samples:

$$X = (x_1, x_2, \cdots, x_n). \tag{1}$$

The sample is independent and identically distributed, thus, a research method for unsupervised learning is to define a $(n \times d)$ matrix whose rows represent samples:

$$X = \left(x_i^T\right)_{i\in[n]}^T. \tag{2}$$

The purpose of unsupervised learning is to discover the different structural information and laws contained in the matrix X.

Unsupervised learning does not pretrain on training samples, there is also no supervision information available, and the feature library of the samples cannot be established. If the classifier continues to accept a large number of edge test samples, it may affect the classification accuracy, resulting in misclassification [19].

3.1.2. Supervised Learning. Supervised learning is a traditional machine learning method, and it utilizes the prior knowledge provided by the system (such as the labeled class information of the sample, pairwise constraint information, and prior probability), learns the known training sample set, adjusts the parameters of the classifier, and establishes a sample learning model, then, the classification of unknown

samples is realized according to the sample model [20]. In supervised learning, the sample set X and the class label of the sample:

$$Y = (y_1, y_2, \cdots, y_n). \quad (3)$$

This is known, where y_i is the class label corresponding to sample x_i, and data pair (x_i, y_i) constitutes the training set of samples needed to construct the learner. Supervised learning, hoping to find the mapping relationship between x and y through the known training set, constructs the required learner accordingly.

Labeled data is often difficult to obtain in supervised learning, and it is necessary to establish a feature library in the training phase of supervised learning, and this will easily lead to the features of the new data may not match the features in the library, resulting in the possibility of misclassification.

3.1.3. Semisupervised Learning. Semisupervised learning is a method between unsupervised learning and supervised learning, the data set used in its learning process usually contains a small amount of labeled information, through these samples of identification information, constraints guide the learning of unknown samples [21].

In semisupervised learning, the entire dataset $X = (x_1, x_2, .., x_n)$ is divided into two parts: known labeled dataset $X = (x_1, x_2, .., x_l)$, corresponding labeled $Y = (y_1, y_2, .., y_l)$, and unknown labeled datasets:

$$X_u = (x_{l+1}, x_{l+2}, \cdots, x_{l+u}). \quad (4)$$

The main content to be studied in semisupervised learning, it is how to comprehensively utilize labeled samples and unlabeled samples.

According to different learning methods, common semisupervised learning algorithms can be classified into the following categories:

(1) *Generative Model Algorithm*. Such algorithms usually use generative models as learners, the probability of dividing unlabeled samples into each class is regarded as a set of missing parameters, then apply the EM (expectation—maximization; expectation maximization) algorithm, estimate feature labels and model parameters. This algorithm is a widely used method in the early days and can be regarded as clustering within the range of a small number of labeled samples

(2) *Based on the Graph Regularization Algorithm*. This kind of algorithm usually adopts the manifold assumption and generally builds a graph based on the training samples and the relationship between them, the connecting line in the figure represents the similarity between samples, then, define the objective function to be optimized, based on the smoothness of the graph as regularization, use the decision function to find the optimal model parameters

(3) *Cotraining Algorithms*. These methods use two or more classifiers and improve classification accuracy through collaboration between classifiers. The classifier trained each time marks the unlabeled samples and selects samples with higher confidence to add to each other's training sets, this is continuously updated until a certain condition is met, thereby, the classification interface is gradually updated

3.2. Cluster Analysis Theory

3.2.1. Definition of Cluster Analysis. A class or cluster is a collection of data objects, the data objects in the same cluster are similar to each other, unlike objects in other clusters. From a machine learning point of view, cluster analysis is a type of supervised learning [22]. Before performing cluster analysis on the data, we do not know how many categories we will eventually divide into, instead, it is clustered according to the similarity of information between the data; finally, the similarity between individuals of the same class is maximized, while the similarity between individuals of different classes is minimized.

3.2.2. Basic Steps of Clustering. The clustering methods adopted for different problems are also different, but they are all based on a specific process frame, as shown in Figure 2, broadly speaking, most clustering methods have four steps: feature selection or feature extraction, clustering algorithm design or selection, cluster confirmation, and result interpretation. It is also a process of transforming useless data into useful knowledge. In a narrow sense, clustering includes the design and selection of clustering algorithms, the process of clustering confirmation, and the interpretation of results.

3.2.3. Classification of Clustering Methods. (1) Partitioning method: this kind of method is easy to describe and simple to implement, and it is also the most studied clustering method, the division method requires a given number of clusters k to be divided into, when classifying, it is first necessary to obtain a set of k initial divisions, then adopt the method of iterative relocation, and improve the quality of partitioning by moving data objects from one cluster to another [23]. The process can be represented as a binary hierarchical tree, the leaf node represents a data object, the middle point indicates that the dataset is split into two distinct classes, or a class is merged from its two subclasses. Density-based methods break through this limitation, the main idea is to give a density threshold, as long as the point in the region is higher than the threshold, it is assigned to the cluster close to it, that is, the purpose of clustering is achieved by finding high-density regions divided by low-density. (2) Grid-based method (grid-based method): grid-based methods first quantify the object space and are divided into a certain number of cells, which are called grids, and the next clustering operation is performed in this quantized grid structure. (3) Model-based method: assuming that a given

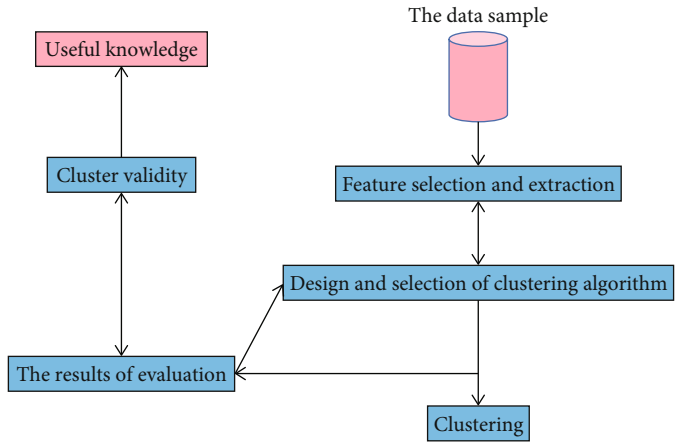

Figure 2: Clustering process diagram.

data distribution has underlying regularities, then, model-based methods try to find out this pattern, find some kind of mathematical model, and fit it to the given data.

3.3. K-Means Algorithm. The K-means algorithm is the most classic division method, K-means is a centroid-based technique, it takes k as a parameter and divides n objects into k clusters, in order to have a high similarity within the class, and the intraclass similarity is lower. The flow of the K-means algorithm is as follows.

Algorithm 1. Input: Parameter k, n data objects.
Output: k clusters.
Step:
(1) Select k points as the initial centroids
(2) repeat
(3) Assign n data objects to the nearest centroids to form k clusters
(4) Recalculate the centroid of each cluster
(5) until the centroid no longer changes.

The K-means algorithm has the advantages of simple use, fast convergence, and low memory overhead, at the same time, there are some shortcomings, such as the performance of the algorithm depends on the initialized k cluster prototypes, and the clustering performance is unstable, sensitive to noise outliers, etc.

3.4. Semisupervised Clustering Algorithm. Semisupervised clustering is a new clustering method, and it combines the characteristics of supervised learning and unsupervised learning, using a small amount of labeled data as supervision information improves the quality of clustering [24]. Depending on how supervised information is used, semisupervised clustering can be divided into three categories:

(1) Such methods use clustering to constrain the search process of clusters using supervisory information and guide the algorithm to obtain good clustering results. Typical algorithms include COP-K-means method, seed-K-means, and constrained-K-means

(2) *Distance-Based Semisupervised Clustering Method.* Such methods utilize identifying data, train a similarity metric that satisfies constraint information or categories, and then use distance-based clustering algorithm for clustering

(3) *Method Based on Constraint and Distance Fusion (Constraint and Distance-Based Semisupervised Clustering Method).* This type of method combines the above two methods to use, one of the typical algorithms is MPC-K-means algorithm

Semisupervised clustering algorithms have been widely used in practical fields, and these include biological information processing, image processing, text classification, and intrusion detection [25]. According to the needs of the subsequent intrusion detection algorithm, the semisupervised clustering algorithm process adopted by the author is as follows.

Algorithm 2. Input: Parameter k, labeled dataset S_l, unlabeled dataset S_u.
Output: k clusters
(1) Use the labeled data in S_l to determine L initial cluster centroids
(2) $\forall x \in S_u$ calculates the minimum distance from each cluster centroid, take the data point corresponding to the maximum value of the minimum distance, as the centroid of the next cluster, record it as the L+1 centroid
(3) $\forall x \in S_u$ calculates its distance from each cluster centroid, assign x to the cluster to which the centroid with the smallest distance belongs, and update the centroid of each cluster
If the cluster centroid is k, repeatedly assigning each data point in S_l and S_u to the cluster to which the cluster centroid with the smallest cluster distance belongs, recalculate the centroid of each cluster, otherwise turn to step (2);
(4) Cluster the k cluster centroids until the cluster centroids no longer change
(5) Output k clusters.

In the initial stage of the above algorithm, the initial cluster center is generated by using the label information of the

data, make the initial cluster center controllable, the robustness of the system is enhanced by the method of gradually generating cluster centers, and the convergence speed and accuracy of the clustering algorithm have been improved.

3.5. Cloud Model-Based Semisupervised Clustering Intrusion Detection Algorithm. Whether it is a K-means clustering algorithm or a semisupervised clustering algorithm, there is a problem of threshold division in the intrusion detection algorithm, the value of the threshold directly affects the detection result, and in practice, it cannot flexibly respond to intrusion situations. Furthermore, the application of general cloud model classifiers in intrusion detection, it is often implemented through an association rule generator, which is slow in processing, considering the properties of network data is not comprehensive, in practical applications, it cannot cope with the complex and changing network environment. Based on the above methods, the author combines the algorithm of semisupervised clustering and the characteristics of cloud model, this aspect does not require thresholding after the initial clustering, extract relevant data directly from a small amount of identification information, build a cloud model classifier, and use a dynamic weighting method in the classification process, and the flexible use of real-time detection data makes the weighting method more reasonable. After a certain data is classified, the data record and other data learn from each other, continuously adjust the cloud model, enhancing its classification ability, the classifier can adapt to the changing network environment.

3.5.1. Relative Proximity of Clouds. Cloud relative closeness is proposed on the basis of cloud model theory, it reflects the degree of similarity between clouds and fully express the randomness and ambiguity of evaluating language concepts, and it is in line with people's subjective feelings and has greater objectivity. Its specific definition is as follows.

Suppose there are two clouds $A1(Ex1, En1, He1)$ and $A2(Ex2, En2, He2)$ in the universe of discourse space U, define $D_{1,2} = |Ex1 - Ex2|$, then, $D_{1,2}$ reflects the relative closeness of the two clouds.

3.5.2. Weighted Intrusion Detection Algorithm. The reverse cloud generator obtains the digital features of the cloud from the real training set, form judgment rules, realize normal modeling, in practice, this method requires a large amount of training data and training time, the cloud digital eigenvalues obtained from the training data does not reflect the actual situation at the time of the invasion, and the calculation of attribute weights in the article is too subjective, at the same time, it is very difficult to determine the threshold value during detection.

The author first uses a semisupervised clustering algorithm to cluster the dataset, then, the results of the clustering are arranged in order of the size of the clusters, at the same time, the normal data clusters and abnormal data clusters are preliminarily screened out according to the tag information, use the data in the cluster to build a normal cloud model and an abnormal cloud model, with the improved 1D inverse cloud generator and X-condition cloud generator, build a cloud model classifier to classify the remaining data objects, at the same time, the classification adopts the method of continuously updating the cloud model and recalculating the weight of each attribute, in order to guide the classification of data. Since the relative closeness of the cloud has greater objectivity, the author refers to this concept for the setting of attribute weights, that is, assuming that the normal cloud is Al in intrusion detection, the abnormal cloud is A2, then, when building a cloud model for each dimension attribute, the size of $D_{1,2}$ reflects the degree of difference between normal clouds and abnormal clouds and the relative importance of this attribute in the classification process. Using this method to weight attributes is in line with people's cognition of the concept of things, the dynamic weighting method can make full use of the implicit information of the data itself, and the weighting method is more scientific.

The steps of the intrusion detection method based on semisupervised clustering are as follows.

Algorithm 3. Input: A dataset S containing nd-dimensional data, $S = S_l \cup S_u$ (labeled dataset S_l, unlabeled dataset S_u).

Output: The data type of data $x \in S_u$ (normal or abnormal).

(1) Use the semisupervised clustering algorithm in 4.4 to cluster the dataset S

(2) Arrange the clustering results in ascending order according to the size of the clusters

(3) Combined with the label information of the data, the initial normal clusters and abnormal clusters are screened out as C_n and C_a, respectively, the rest of the data is allocated into C_r

(4) For each dimension of data in C_n, the corresponding cloud digital eigenvalue $(Ex1_i, En1_i, He1_i), i = 1, \cdots, d$ is obtained by using the reverse cloud generator

(5) For each dimension of data in C_a, use the reverse cloud generator to obtain the corresponding cloud digital eigenvalue $(Ex2_i, En2_i, He2_i), i = 1, \cdots, d$

(6) Use formula (5) to calculatse the weight of each attribute

$$w_i = \frac{|Ex1_i - Ex2_i|}{\sum_{j=1}^{d} |Ex1_j - Ex2_j|}, \tag{5}$$

(7) Take a data object x from C_r in turn, according to the X-condition forward cloud generator using the formula, the abnormal and normal cloud classification models are calculated:

$$\mu_j = \sum_{i=1}^{d} w_i \cdot \exp\left[\frac{-(x - Exj_i)}{2 \cdot Enj_i}\right], j = 1, 2. \tag{6}$$

If $\mu_1 > \mu_2$ then x belongs to the normal class, assign it to C_n, after returning to step (4) to update the normal cloud model, go to step (6) to recalculate the weight of each attribute, otherwise, assign x to C_a and return to step (5) after updating the abnormal cloud model, then, go to step (6) to

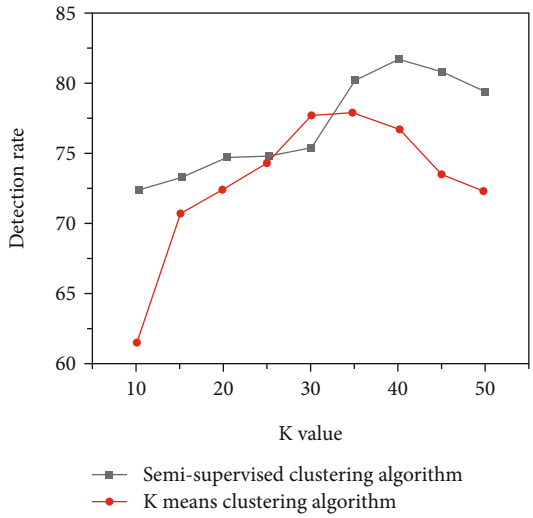

Figure 3: Comparison of detection rate detection results.

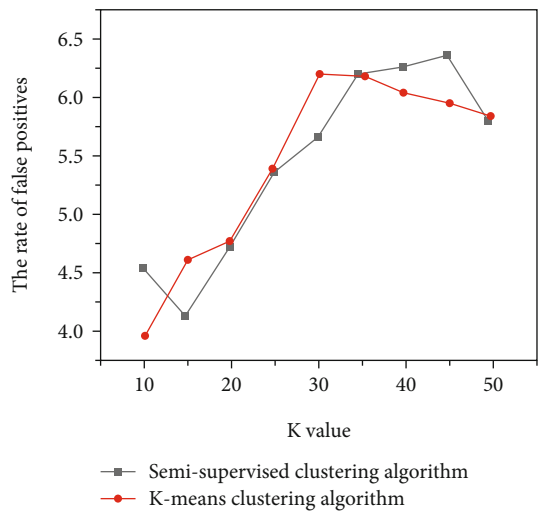

Figure 4: Comparison of false positive rate detection results.

recalculate the weight of each attribute until all data classification ends.

4. Results and Analysis

4.1. Semisupervised Clustering Algorithm Experimental Data Selection and Preprocessing. Since the original dataset is too large, we select some representative data for experiments, a subset U was selected for testing from "kddcup.newtestdata_10__percent_corrected," and selected 500 records as identification data records, 10000 records as test data, among them, 758 were DoS attacks, 15 were R2L attacks, 42 were U2R attacks, and 92 were probe attacks.

In the experiment, we use the experimental platform based on MATLAB, the KDDCUP99 dataset contains symbolic data attributes, which cannot be recognized by MATLAB; therefore, it is necessary to renumber the attribute value of the symbol type and use the natural number set to renumber the seven values, taking protocol_type as an example, tcp, udp, and icmp are replaced by natural numbers 1, 2, and 3, respectively. And so on, the original data will become a numeric type. There are two types of numerical variables, one is a continuous attribute characteristic variable, and the other is a discrete attribute characteristic variable. For continuous attribute feature variables, attribute characteristics of different attributes may have different metrics, if before experimenting, if the data is not preprocessed, there may be a problem of large numbers eating decimals, as a result, the attribute characteristics of some values are masked, thereby affecting the experimental results, therefore, in the process of data preprocessing, it is necessary to normalize and normalize attribute values.

4.1.1. Standardization.

$$x_{ij} = \frac{x_{ij} - m_j}{S_j}, \quad (i = 1, \cdots, n\,; j = 1, \cdots, r). \tag{7}$$

Among

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad s_j = \sqrt{\frac{1}{n-1}\left(x_{ij} - m_j\right)^2}. \tag{8}$$

4.1.2. Normalization.

$$x_{ij}{}' = \frac{x_{ij} - \min\left(x_{ij}\right)}{\max\left(x_{ij}\right) - \min\left(x_{ij}\right)}. \tag{9}$$

Among $i = 1, \cdots, n\,; j = 1, \cdots, r$.

4.2. Simulation Experiment and Result Analysis. The experiments were run on a machine with CPU 2.2 GHz, 2.00 GB, Microsoft Windows XP, and adopt MATLAB7.8.0 to realize. In order to evaluate the performance of intrusion detection methods, the experiment adopts detection rate and false alarm rate as the metrics of algorithm performance. Its definition is as follows:

$$\text{Detection rate} = \frac{\text{number of detected attacks}}{\text{total number of attacks}}, \qquad \text{False positive rate} = \frac{\text{the number of normal samples that were falsely reported as intrusions}}{\text{the number of normal samples}}. \tag{10}$$

Table 1: Experimental test data table.

Test data	Quantity	DoS (%)	R2L (%)	U2R (%)	Probe (%)
Dataset 1	5844	3.56	0.87	0.55	1.18
Dataset 2	5836	3.72	0.87	0.46	1.18
Dataset 3	5737	3.78	0.89	0.45	1.29

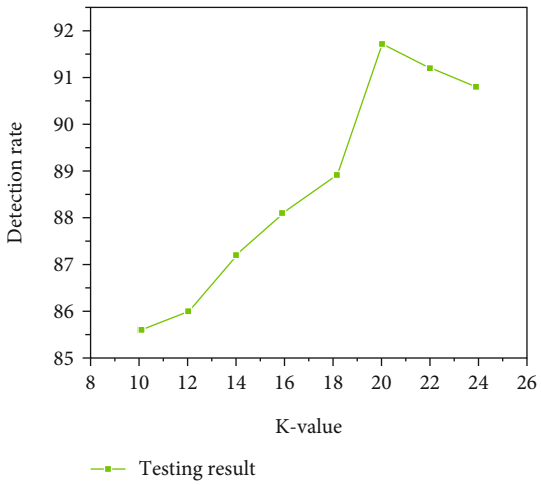

Figure 5: Detection rate under different K values.

For the above 10,000 records, the K-means clustering algorithm and the semisupervised clustering algorithm described by the author were used, respectively, intrusion detection data tests were performed, Figures 3 and 4 show that under different K values, average detection rate and false positive rate under the K-means clustering algorithm and the authors' semisupervised clustering algorithm.

By comparison, it can be seen that the detection rate of the author's semisupervised clustering algorithm is higher than that of the K-means clustering algorithm, and the false alarm rate and K-means algorithms have also been improved. Therefore, the author's semisupervised algorithm enhances the stability of the system, and the performance of the K-means algorithm is improved to a certain extent.

4.3. Experimental Data Selection of Semisupervised Clustering Intrusion Detection Algorithm. Experimental simulation of semisupervised clustering intrusion detection algorithm for cloud model, three sets of data sets were selected for testing from "kddcup.newtestdata_10_percent_corrected," at the same time, 500 pieces of data are selected as identification data records. The test data types and distributions are shown in Table 1.

Among them, the attacks of Do S are smurf and neptune; the attacks of R2L are guess passwd; U2R's attack is buffer_overflow, land module, perl, and rootkit; and probe's attack is port sweep.

4.4. Simulation Experiment and Result Analysis. Different K values were selected to test the above three sets of data, respectively, take their average as the detection result. Figures 5 and 6 show the detection results of the cloud

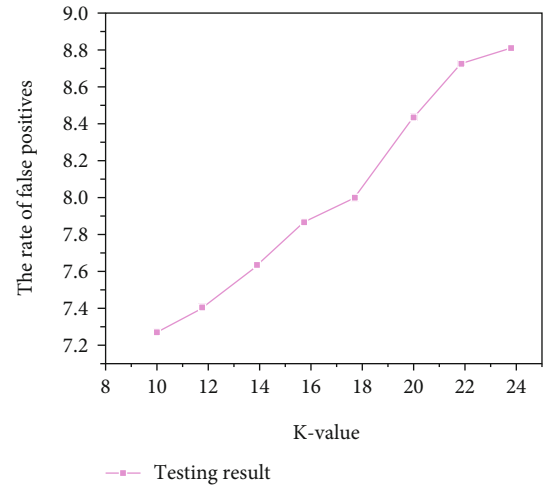

Figure 6: False alarm rate under different K values.

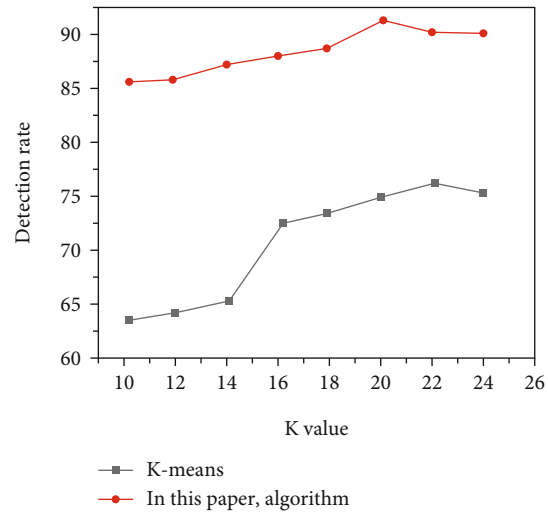

Figure 7: Comparison of detection rate results under different K values.

model semisupervised clustering algorithm under different K values.

From the experimental results in the figure, it can be seen that, when the value of K gradually increases, the false alarm rate also increases; however, when K is 20, the detection rate is maximized. From this, it can be known that when K is 20, the semisupervised clustering algorithm based on cloud model can obtain better intrusion detection effect, its detection rate reaches 91.76%, and the false alarm rate is 8.54%. In the cloud model-based semisupervised clustering algorithm, the detection rate and false alarm rate when K takes different values are compared with the K-means algorithm, as shown in Figures 7 and 8.

As can be seen from the above figure, in the case of different values of K, the author's algorithm is significantly higher than the K-means algorithm in terms of detection rate, the false positive rate is slightly higher than that of K-means, but this false positive rate is within an acceptable range.

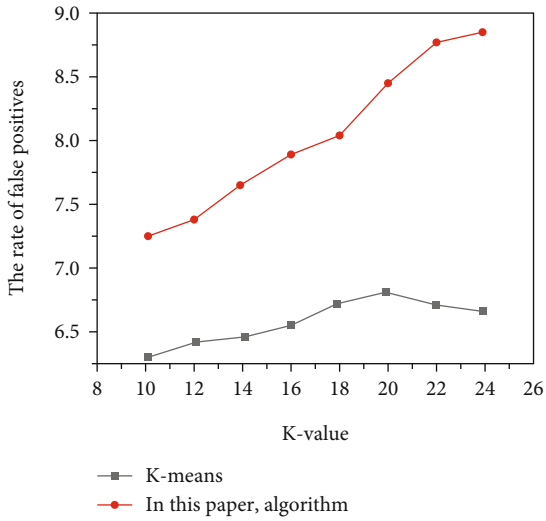

Figure 8: Comparison of false alarm rate results under different K values.

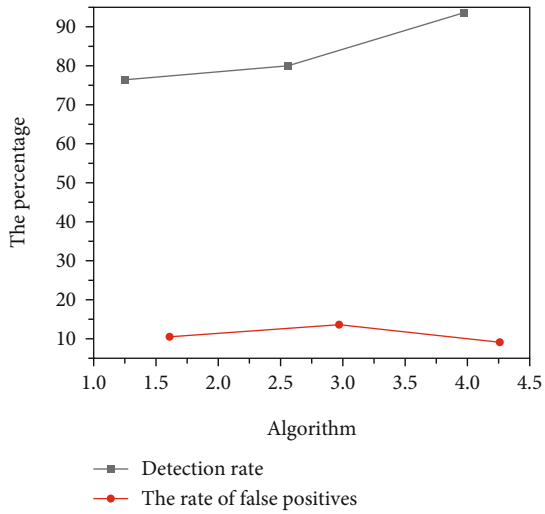

Figure 9: Comparison of test results.

The detection results of the cloud model-based semisupervised clustering algorithm, and the comparison results of the general clustering algorithm and the general cloud model classifier, are shown in Figure 9.

The above figure shows the comparison between the detection results of the author's algorithm and several other algorithms, by comparison, it can be found that the detection rate of the author's algorithm is significantly higher than the other two algorithms, and the false positive rate is slightly higher than that of K-means, the false positive rate is lower than that of the other algorithm, proving the superior performance of our algorithm.

5. Conclusion

The author proposes a research on computer network confidential information security based on big data clustering algorithm, intrusion detection data has the characteristics of high-dimensional attributes, currently, cloud model classifiers can only handle one-dimensional and two-dimensional data, and the traditional intrusion detection method based on semisupervised learning relies heavily on prior knowledge. In order to solve the above problems, the author proposes a semisupervised clustering intrusion detection algorithm based on cloud model, a new cloud model classifier is constructed, and make full use of a small number of labeled samples and unlabeled samples to guide the classification of data. For the semisupervised clustering intrusion detection algorithm proposed by the author, the simulation experiment of the intrusion detection method based on cloud model semisupervised clustering is done. Analyzing the author's intrusion detection method and comparing it with several other common intrusion detection algorithms, the results show that the method proposed by the author improves the performance of the intrusion detection system, the detection accuracy is improved, the false alarm rate is reduced, and the robustness of the system is enhanced. In the author's algorithm, since there are few labeled samples, it provides correct guidance for the initial formation of normal and abnormal cloud models. The dynamic weighting method is used to solve the problem that high-level data is difficult to process and enable the data to learn from each other and gradually form a relatively stable cloud model, and over-reliance on prior knowledge of the data is avoided. In contrast, compared with the general clustering algorithm, the author's algorithm has a great improvement in performance, to a certain extent, solve some problems existing in the current intrusion detection, however, there are still some problems, such as the false positive rate is still high, and it is sensitive to some data with special distribution and cannot obtain a good classification effect, in future research, further improvements to the algorithm are still needed.

Data Availability

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

Conflicts of Interest

The author do not have any possible conflicts of interest.

References

[1] Y. Guo and S. Yan, "Research on hospital computer network topology based on complex network theory," *Basic & Clinical Pharmacology & Toxicology*, vol. 118, 1, pp. 114–114, 2016.

[2] Y. Chen, H. Zhang, L. Liu, X. Chen, and J. Xie, "Fuzzy risk analysis based on the ranking of generalized trapezoidal fuzzy numbers," *Applied Intelligence*, vol. 26, no. 1, pp. 1–11, 2007.

[3] Y. Lun, X. Zhang, and J. Q. Zhao, "Research on prediction method of anti viral polymer drug efficacy based on neural network algorithm," *Basic & Clinical Pharmacology & Toxicology*, vol. 119, no. 4, pp. 46–46, 2016.

[4] G. Wang, X. Tian, J. Geng, and B. Guo, "A knowledge accumulation approach based on bilayer social wiki network for computer-aided process innovation," *International Journal of Production Research*, vol. 53, no. 8, pp. 2365–2382, 2015.

[5] R. Sharma, V. Vashisht, and U. Singh, "Woatca: a secure and energy aware scheme based on whale optimisation in clustered wireless sensor networks," *IET Communications*, vol. 14, no. 8, pp. 1199–1208, 2020.

[6] Q. Ding, Y. Wu, and W. Liu, "Molecular mechanism of reproductive toxicity induced by Tripterygium wilfordii based on network pharmacology," *Medicine*, vol. 100, no. 27, article e26197, 2021.

[7] Y. J. Liang, C. Ren, H. Y. Wang, Y. B. Huang, and Z. T. Zheng, "Research on soil moisture inversion method based on ga-bp neural network model," *International Journal of Remote Sensing*, vol. 40, no. 5-6, pp. 2087–2103, 2019.

[8] W. Xiong, C. M. Cheung, P. Sander, and A. Joneja, "Rationalizing architectural surfaces based on clustering of joints," *IEEE transactions on visualization and computer graphics*, vol. 1, p. 1, 2021.

[9] D. H. Kim, K. H. Choi, K. J. Li, and Y. S. Lee, "Performance of vehicle speed estimation using wireless sensor networks: a region-based approach," *Journal of Supercomputing*, vol. 71, no. 6, pp. 2101–2120, 2015.

[10] A. A. L.-S. Ye and S. Lu, "Quantum image steganography and steganalysis based on lsqu-blocks image information concealing algorithm," *International Journal of Theoretical Physics*, vol. 55, no. 8, pp. 3722–3736, 2016.

[11] P. Qian, T. Shang, Y. Gao, and G. Ding, "Research on dynamic handover decision algorithm based on fuzzy logic control in mobile fso networks," *Photonic Network Communications*, vol. 41, no. 2, pp. 136–147, 2021.

[12] B. Yao, L. Wang, and S. Liu, "Research on ocean government data extraction and clustering based on xml document similarity technology," *Journal of coastal research*, vol. 98, no. sp1, p. 259, 2019.

[13] Y. Meng, Y. Chen, F. Zhu, and E. Tian, "The integration of marine biodiversity information resources based on big data technology," *Journal of coastal research*, vol. 103, no. sp1, p. 806, 2020.

[14] D. T. Monteith, P. A. Henrys, C. D. Evans, I. Malcolm, E. M. Shilland, and M. G. Pereira, "Spatial controls on dissolved organic carbon in upland waters inferred from a simple statistical model," *Biogeochemistry*, vol. 123, no. 3, pp. 363–377, 2015.

[15] B. Kirubakaran and M. Ilangkumaran, "Selection of optimum maintenance strategy based on fahp integrated with gra–topsis," *Annals of Operations Research*, vol. 245, no. 1-2, pp. 285–313, 2016.

[16] J. Ge and J. Liu, "Security assessment algorithm of navigation control system based on big data," *Journal of coastal research*, vol. 93, no. sp1, p. 1026, 2019.

[17] Z. Cui and J. Yang, "Research on ocean big data service technology in distributed network environment," *Journal of coastal research*, vol. 98, no. sp1, p. 141, 2019.

[18] W. Zhou and S. Yu, "Research on the communication method of mobile network shadow fading based on interference alignment algorithm," *Journal of Supeuting*, vol. 72, no. 7, pp. 2891–2909, 2016.

[19] J. S. Teh, A. Samsudin, and A. Akhavan, "Parallel chaotic hash function based on the shuffle-exchange network," *Nonlinear Dynamics*, vol. 81, no. 3, pp. 1067–1079, 2015.

[20] C. Lu, "Research on optimization of computer network quality of service based on improved red algorithm," *Revista de la Facultad de Ingenieria*, vol. 32, no. 4, pp. 321–328, 2017.

[21] T. Steiner, R. Verborgh, J. Gabarro, E. Mannens, and R. Walle, "Clustering media items stemming from multiple social networks," *Computer Journal*, vol. 58, no. 9, pp. 1861–1875, 2015.

[22] I. Hababeh, I. Khalil, and A. Khreishah, "Designing high performance web-based computing services to promote telemedicine database management system," *IEEE Transactions on Services Computing*, vol. 8, no. 1, pp. 47–64, 2015.

[23] Y. Duan, R. Yang, and S. Duan, "Overall layout and security measures of campus wireless local area networks," in *The International Conference on Cyber Security Intelligence and Analytics CSIA 2020: Cyber Security Intelligence and Analytics*, pp. 25–32, Springer, Cham, 2020.

[24] P. Xia, "Data security risk and preventive measures of virtual cloud server based on cloud computing," in *The International Conference on Cyber Security Intelligence and Analytics CSIA 2020: Cyber Security Intelligence and Analytics*, pp. 40–45, Springer, Cham, 2020.

[25] Z. Zheng, "Information security risk assessment based on cloud computing and bp neural network," in *The International Conference on Cyber Security Intelligence and Analytics CSIA 2020: Cyber Security Intelligence and Analytics*, pp. 85–91, Springer, Cham, 2020.