Research on Anomaly Detection of Distributed Intelligent Teaching System Based on Cloud Computing

Fayue Zheng(✉), Lei Ma, Hongxue Yang, and Leiguang Liu

Beijing Polytechnic, Beijing 100016, China
fayue@126.com

Abstract. The traditional anomaly detection method of intelligent teaching system has some problems, such as poor accuracy and response efficiency. Therefore, this paper proposes a distributed anomaly detection method of intelligent teaching system based on cloud computing. Collect the abnormal data of distributed intelligent teaching system through cloud computing method, calculate the local reachable density according to Gaussian distribution function, build a data management model, and use distributed technology to locate and manage the abnormal area of teaching data, so as to achieve the goal of data detection and identification. The experimental results show that this method can effectively improve the recall rate of anomaly detection data in intelligent teaching system, and the response efficiency has been effectively improved.

Keywords: Cloud computing · Distributed · Intelligent teaching · Anomaly detection

1 Introduction

With the progress of science and technology and the rapid development of network technology, the information industry and its application have been greatly developed. Enterprises and institutions such as government, finance, education and individual users are more and more dependent on the network. At the same time, it also brings hidden dangers of information security. How to ensure the security of network and information system has become a highly valued problem [1–3]. As an active security protection technology, intrusion detection can detect and identify external or internal abnormal activities or intrusion behaviors, such as malicious use or destruction of computer and network resources, unauthorized access of internal users, etc. it has become a useful supplement to the traditional computer security technology, It is a new hotspot in the field of network security.

Reference [4] proposes a distributed intelligent teaching system anomaly detection method based on deep learning, which obtains the data transmitted by the intelligent teaching system through the data mining method, constructs the network training model

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 453, pp. 709–723, 2022.
https://doi.org/10.1007/978-3-031-21161-4_54

according to the deep learning method, and realizes the anomaly detection of the intelligent teaching system. This method can improve the monitoring response time, but the anomaly monitoring recall rate is poor. Reference [5] proposes a teaching system anomaly detection method based on ZigBee technology, which uses sensors to obtain abnormal data of the teaching system, and uses ZigBee technology to repair abnormal problems in the teaching system. This method can improve the accuracy of anomaly monitoring, but it is time-consuming. Reference [6] proposes a teaching system anomaly detection method based on blockchain technology. The clustering center of teaching system anomaly is obtained through data clustering method, and the blockchain technology is used to detect the learning system anomaly. This method can improve the anomaly detection effect, but the response speed is poor.

To solve the above problems, this paper proposes an anomaly detection method for distributed intelligent teaching system based on cloud computing. Collect the abnormal data of the distributed intelligent teaching system through cloud computing method, calculate the local reachable density according to the Gaussian distribution function, build a data management model, use the distributed technology to locate and manage the abnormal area of teaching data, achieve the goal of data detection and identification, and effectively improve the accuracy of abnormal detection of the intelligent teaching system.

2 Anomaly Detection of Distributed Intelligent Teaching System

2.1 Recognition of Abnormal Information in Intelligent Teaching System

The data that intrusion detection needs to analyze is collectively referred to as events. It can be packets in the network or information obtained from host system logs and other ways. The purpose of event generator is to obtain events from the whole computing environment and provide this event to other parts of the system in a specific format [7]. The event analyzer analyzes the data to determine whether it is violation, anomaly or intrusion, and converts the judgment result into warning information. The response unit responds according to the warning information. It can make a strong response such as

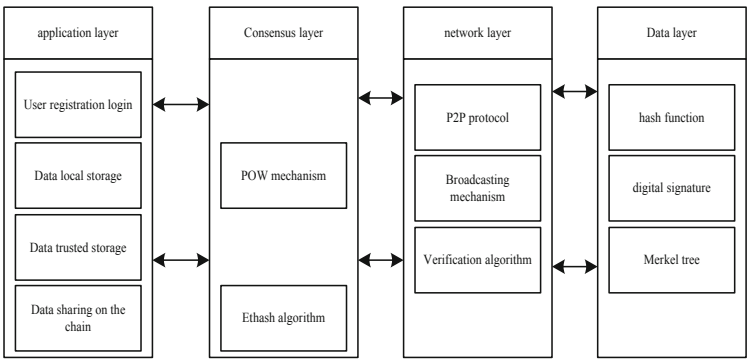

Fig. 1. The structure design of the information database of the intelligent teaching system

cutting off the connection and changing the file attributes, or it can be just a simple alarm. It is an active weapon in intrusion detection. The online learning behavior data storage and sharing model based on blockchain technology optimizes the information database structure of intelligent teaching system, as shown in the Fig. 1:

The abnormal event database based on cloud computing is a place to store various intermediate and final data. It receives data from the event generator or analyzer and saves it for a long time [8]. It can be a complex database or a simple text file. In this model, the first three appear in the form of program, while the last one often appears in the form of file or data flow. The specific model is as follows (Fig. 2):

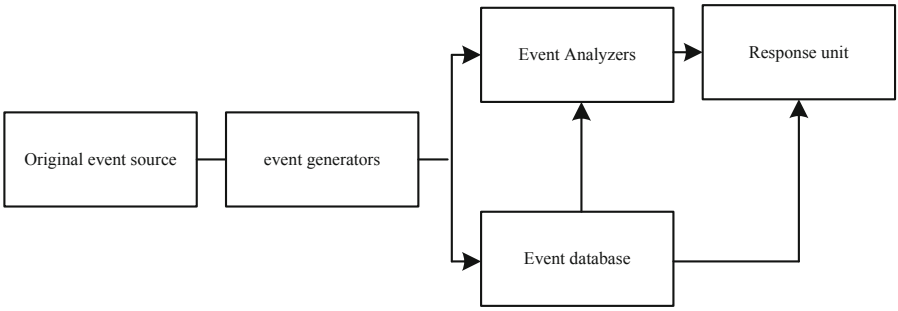

Fig. 2. Intrusion detection model based on cloud computing

The anomaly detection system of intelligent teaching system attempts to establish a feature prototype corresponding to "normal activities", and then mark all behaviors that are "very different" from the established feature prototype as anomalies. Deep learning is a meaningful learning process, which requires the joint participation of teachers and students. Teachers play the role of guides. Students actively participate, experience success and grow around the challenging learning theme [9]. Students not only master the core knowledge, essence and thinking methods of the subject, but also form a high learning motivation, positive learning attitude and correct values. The table is a comparison of cognitive goals between deep learning and shallow learning (Table 1).

Table 1. Comparison of cognitive objectives between deep learning and shallow learning

Learning type	Target hierarchy	Connotation
Deep learning	Application	Apply the learned skills to the new situation
	Analysis	Decompose the material into elements and clarify the relationship between elements and the whole
	Evaluate	Make value judgment on the learned knowledge according to the criteria

(*continued*)

Table 1. (*continued*)

Learning type	Target hierarchy	Connotation
	Create	Integrate all elements into a consistent whole to form a new model or structure
Shallow learning	Memory	Extracting relevant information from long-term and short-term memory
	Understand	Understanding the meaning of knowledge from teaching information

It is assumed that all abnormal behaviors are different from normal behaviors. If the trajectory of the normal behavior of the system can be established, all system states different from the normal trajectory can be regarded as suspicious attempts in theory. As shown in the Fig. 3:

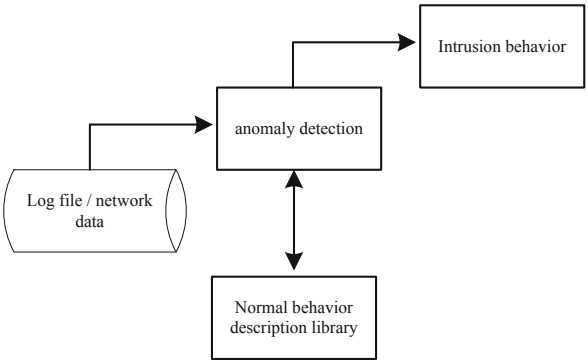

Fig. 3. Cloud computing anomaly detection model

In the process of processing multidimensional data, the isolated forest algorithm uses the method of randomly selecting attributes to build a tree, and finally integrates the results of each tree to judge the anomaly, while ignoring the problem that each instance in the multidimensional data has different anomaly degrees for the randomly selected attributes. Therefore, in general, the abnormal score detected only by random attributes will still be inaccurate, which needs to be further improved.

2.2 The Abnormal Area Location Algorithm of the Intelligent Teaching System

When calculating the learning credit of online learning platform learners, it is necessary to obtain the variable values involved in the relevant formulas first. The background database will collect the data required for anomaly detection while obtaining the source data [10]. Referring to the relevant data of anomaly detection, on the one hand, modify the relevant record data of abnormal learning behavior, on the other hand, deduct the

corresponding anomaly points for the calculated five-dimensional evaluation values, and then obtain the updated five-dimensional evaluation values Pa, Po, Re, Qu and Ln. After calculating the relevant weights, you can calculate the learning credit of each learner participating in the platform learning activities. The calculation formula is:

$$SC = Pa'\kappa_1 + Po'\kappa_2 + Re'\kappa_3 + Qu'\kappa_4 + Ln'\kappa_5 \tag{1}$$

Among them, κ_1, κ_2, κ_3, κ_4, κ_5 are the weights of the five-dimensional evaluation index of learning credit, namely the weight of the five dimensions of Pa, Po, Re, Qu, and Ln relative to the weight of learning credit SC. The statistics-based method is one of the simplest and basic methods in anomaly detection, and its principle is also very easy to understand [11]. This method requires that the data set e must present a known distribution or meet certain laws, and then find the laws that do not meet the requirements. Point. The most common known distribution is the Gaussian distribution. The following uses it as an example to briefly describe the idea of anomaly detection based on statistics. According to the original data set, find the expected x and variance μ to determine the Gaussian distribution function, as

$$f(x) = \left\| \frac{1}{\sqrt{3\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} - SC(x-\mu) \right\| - e \tag{2}$$

When $e < \sigma$ or $e > \sigma$, we find that the probability of its appearance is very small, so we consider the outliers in this interval. This method has some advantages, but there are also disadvantages. In terms of advantages, the outliers obtained by this method have high credibility, and the outliers are intuitively visible. However, this method is limited to conform to a certain distribution [12]. If it does not conform to known distributions and laws, this method is not suitable. Secondly, this method is more difficult in parameter selection, and because outliers are also involved in the construction of the model, if there are too many outliers, it will have a greater impact on the selection of model parameters. Furthermore, this method is not suitable for the detection of high-dimensional data. The basic idea of the distance-based anomaly detection method is to find the distances of X_i points closest to the current point and sum them. The specified distance and the largest Y_i points are the abnormal points. When measuring the distance of anomalous points, it usually refers to Euclidean distance, manhattan distance or Mahalanobis distance. The calculation formulas are as follows:

$$\begin{cases} \text{dis}_{ED} = f(x) + \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} - SC \\ \text{dis}_{MHD} = f(x) - \sum_{i=1}^{n} |X_i - Y_i| + SC \end{cases} \tag{3}$$

The distance anomaly detection algorithm is k-nearest neighbor algorithm (kNN). The abnormal point (a_i, b_i) can be obtained by finding the largest n points with the average distance of k nearest neighbors [13]. The formula for the average distance of k

nearest neighbors of point P is as follows:

$$\mathrm{D}(p) = \frac{p - \sum_{q_i \in N_k(p)} dis(a_i, b_i)}{|\mathrm{dis}_{ED} - \mathrm{dis}_{MHD}|} - N_k(x) \tag{4}$$

Among them, $N_k(x)$ represents k-nearest neighbor, that is, the set of objects whose distance from point p does not exceed k-nearest neighbor. Compared with statistics-based detection methods, distance-based anomaly detection methods can handle multi-dimensional data, but there are also certain problems: first, the choice of distance calculation method and the selection of parameter $\mathrm{lrd}_k(p)$ are difficult; second, it cannot distinguish local outliers. Point. In order to solve the problem that the distance-based anomaly detection method cannot accurately distinguish the local outliers, a density-based outlier method is proposed. The most representative one is the LOF algorithm. The main idea is to judge the abnormal situation by comparing the density of each point with its neighboring points. The smaller the density, the greater the possibility of anomaly. This density is measured by the local outlier factor. The local outlier factor of point p in the LOF algorithm is expressed as

$$LOF_k(p) = \mathrm{D}(p) - \frac{\sum_{o \in N_k(p)}^{\mathrm{lrd}_k(o)} 1/\mathrm{lrd}_k(p)}{|N_k(x) - f(x)|} - SC \tag{5}$$

Among them, $N_k(p)$ is the k distance field, that is, all points within the k distance of point p; $\mathrm{lrd}(p)$ refers to the local reachable density, expressed as

$$\mathrm{lrd}_k(p) = 1/\frac{\sum_{o \in N_k(p)}^{\text{reach - dist}_k(p,o)}}{|N_k(p)|} \tag{6}$$

Among them, reach-distr (pO) represents the reachable distance, that is, the kth reachable distance from o to p. Taking into account the controllability of the data, referring to the existing data collection methods and common behaviors of learners, data collection is performed on several behavior indicators of the evaluation model through Web Server log analysis technology and embedded point technology. Next, the following will focus on introducing the web-side data, online operation data and discussing the collection process of interactive data. The corresponding detailed indicators are shown in the Table 2.

Combined with the weight of refined behavior index elements and standardized learning behavior data, the weight is calculated by structural equation model. The figure shows various learning services provided by the application layer module for users. The main functions include registration, online learning behavior acquisition, learning credit data storage and learning credit data sharing (Fig. 4).

Judge the results of anomaly detection experiments from four aspects. Further judge the accuracy of each deep learning network in the migration model and the newly-built U-CONVLSTM model to learn the normal behavior characteristics of the student; then use the loss value obtained in the training model for each test segment of the test set

Table 2. Correspondence table of data collection types and data collection items

Data collection type	Data collection item
Web side data	IP address, login time, course start time, deadline, study time, number of videos watched, course study days
Online operation data	Click count, collection count, job score, time consumption after video viewing
Discuss interactive data	Original sharing and forwarding times; Times of teacher-student interaction

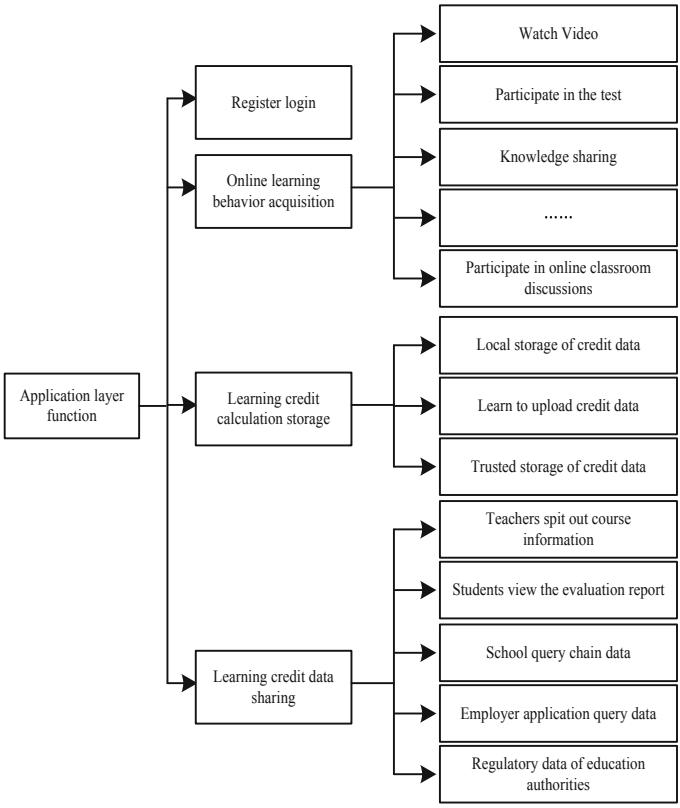

Fig. 4. Application layer intelligent detection function design diagram

to determine at which moment the student is Abnormal behavior occurs; after the loss value is converted into an abnormal score, it can be judged whether the student has abnormal behavior; finally, the test results of all test set data in the migration model and the U-shaped model are counted. In summary, it is concluded that the learning effect and detection effect of the new U-shaped autoencoder is better than the migration model (Fig. 5).

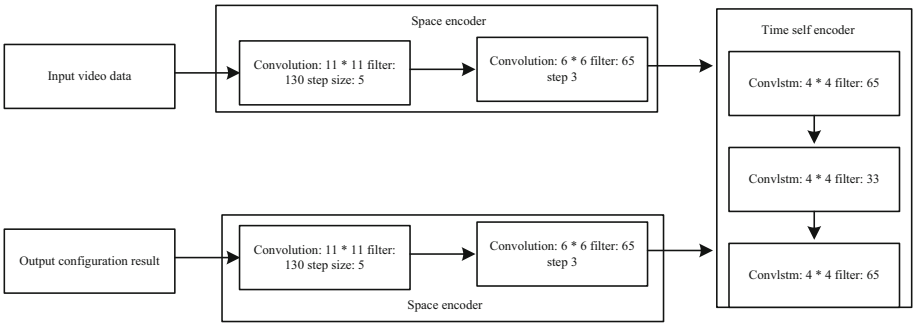

Fig. 5. The learning effect and detection effect are better than the migration model

Unbalanced data means that the number of positive samples is far less than the number of negative samples. Due to this imbalance in number, a more reasonable method needs to be found to analyze these data. Therefore, the analysis method of accuracy, which does not distinguish the importance of each class, is obviously not suitable for analyzing unbalanced data sets. In unbalanced data sets, a small number of analogies and a large number of classes have more research value. Therefore, in binary classification, rare classes are usually marked as positive classes, while most classes are marked as negative classes. The following table shows the confusion matrix of the number of positive and negative samples correctly predicted and incorrectly predicted by the model (Table 3).

Table 3. Confusion matrix

Real class	Prediction class	
	Positive example (+)	Counterexample (-)
Positive example (+)	f + + (AY)	f + + (IN)
Counterexample (-)	f- + (IY)	f--(AN)

The detection results of isolated forest algorithm are fuzzy processed by using the idea of fuzzy logic. The idea of the so-called fuzzy logic is based on the theory of fuzzy mathematics and fuzzy relationship synthesis principle, and uses scientific means to accurately describe some things with fuzzy conceptual boundaries and difficult to quantify. Therefore, it can be detected from multiple angles to ensure the accuracy and effectiveness of anomaly detection results.

2.3 Implementation of Anomaly Detection in a Distributed Intelligent Teaching System

According to the analysis of business requirements, this network traffic security system consists of four subsystems: domain name anomaly detection, forum access monitoring, unhealthy information release traceability, and network abnormal behavior detection.

The functional structure diagram of the entire system is shown in the figure. Collecting and reading network traffic logs and other information are their common functions, and different subsystems will process different contents differently to obtain the required results (Fig. 6).

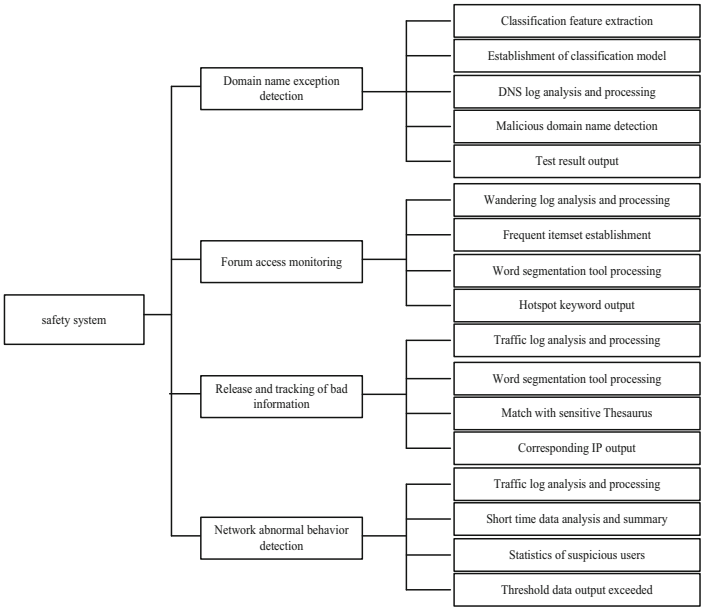

Fig. 6. Functional structure of network traffic security system

The malicious domain name detection module first preprocesses the network traffic log data collected by the traffic collector, extracts the characteristic information that needs to be verified, and stores it on the ES, and then uses the previously established classification model to read the data for classification, and calculate whether these domain names are Malicious domain name, output the malicious domain name and the IP information associated with it, stored in the cluster, and can be downloaded to the machine for use (Fig. 7).

Malicious domain names and normal domain names can be judged based on many factors. These judgment factors are human experience. They are expressed as features, and data feature identification and detection are performed to ensure detection accuracy and efficiency.

3 Analysis of Experimental Results

In order to verify the effectiveness of the algorithm proposed in this paper, this paper uses a real data set in the UC1 data set to conduct a comparative experiment. The processing time is compared and analyzed, and the RoC of each algorithm is drawn. For the method in this paper, the commonly used setting $k = 10$ is used in the experiment. This domain

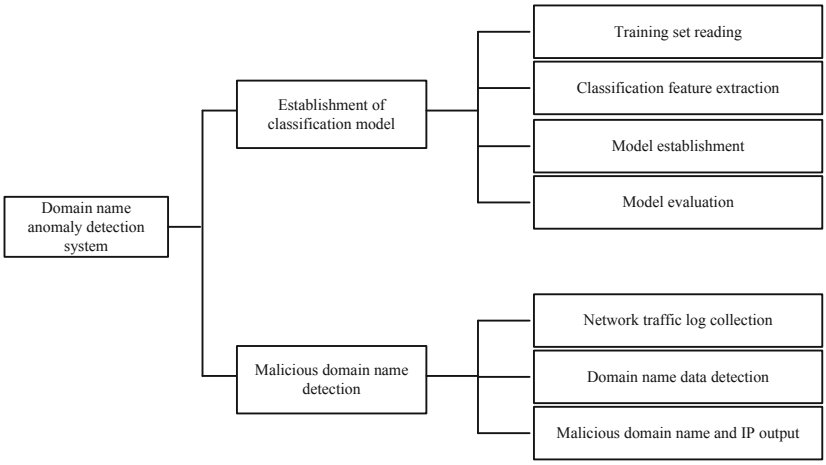

Fig. 7. Functional structure diagram of domain name anomaly detection system

name anomaly detection system is a subsystem of the network traffic security system. The network traffic security system needs to be built on the company's big data cluster. The big data cluster built to test the system functions this time consists of 3 nodes, using the company's CAS The platform creates three virtual machines to build a domain name anomaly detection system. Because the system is developed in Python language, in order to ensure the existence of Python dependent libraries, anaconda2 will be installed, and its specific detailed configuration information will be installed. Use a university's 2019 grades in 2020 and 2020 and 2019 grades in 2019 and 2020 into several groups as data sets (20,000 data), and take the top 1.5% of the larger T(x) data for false detection rate statistics, The parameters are as shown in the Table 4:

Table 4. Experimental parameter setting table

Experiment No	1	2	3
K-means	1	0	0
PFT-OI A	0.7	0.3	0.3
Paper method	0.5	0.4	0.4

The experimental data adopts the teaching evaluation data of students in a university. The original data is composed of four fields. The clustering effect of abnormal detection distance of teaching information is as follows (Fig. 8):

The traditional k-means clustering method has a higher false detection rate for anomaly detection. Compared with the traditional k-means clustering method, the false detection rate of this method is significantly lower, especially when $a = 0.4$, $\beta = 0.3$, $y = 0.3$, the false detection rate reaches a low level. Based on this, the false detection rate

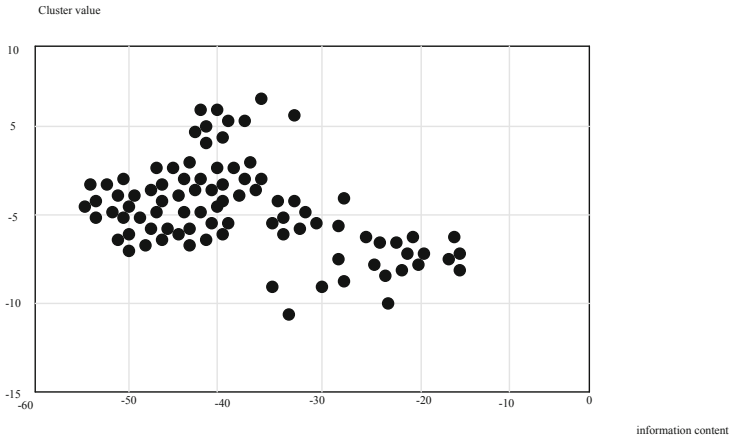

Fig. 8. Distance clustering effect of abnormal detection of teaching information

of data detection under different methods is compared and recorded, and the details are as follows (Fig. 9):

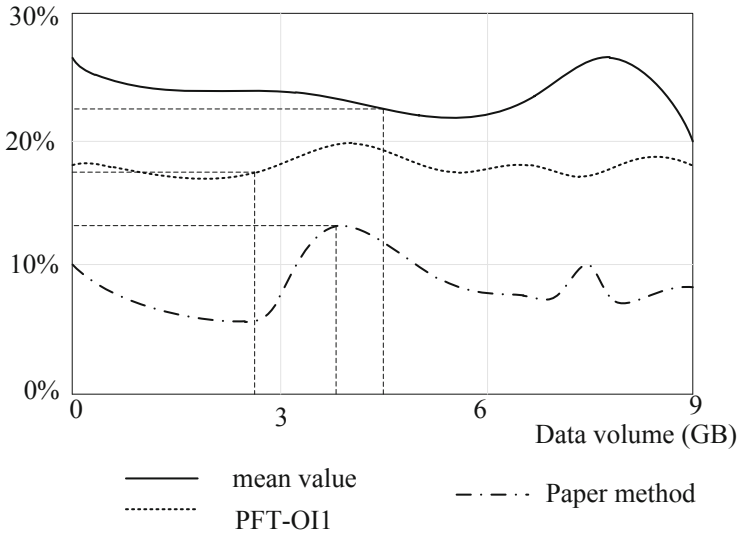

Fig. 9. Statistical chart of system data false detection rate

Using the same data set, it can be seen from the above two experiments that the proposed method has a lower false detection rate compared with the traditional method, which greatly improves the accuracy of anomaly detection. After the establishment of the classification model, the indispensable evaluation of the effect of the classification model shall be carried out. The evaluation sub module mainly uses manual verification to statistically calculate the accuracy rate and recall rate of the classification model. The

main content of its function test is to verify whether the accuracy rate and recall rate are within the expected range to determine whether the classification model meets the requirements. The design of its test cases is shown in the Table 5.

Table 5. Functional test cases of the classification model establishment module

Case number	2.0
Function description	Read the collected feature data, train the classification model, read the domain name data to be tested, and judge whether the output result is correct
Design purpose	Verify whether the classification model can be trained correctly to ensure the correctness of the domain name detection classification model
Preconditions	The python environment configuration is normal, and the server has the characteristic data collected by the training model and the test domain name data needed to verify the rationality of the model
Use case design	Using the collected features to train the classification model, the correctness of the trained model is verified by the test data
Expected results	The program can run normally, and the training can get the classification model, which can classify the domain name
Test result	Generate the classification model correctly and output the test results normally
Test status	Adopt

Through the statistics of 7000 DNS log data used in the test, a total of 173 malicious domain names are obtained. According to the classification models trained on training sets of different sizes, the statistics of the detection output results of these log data are compared and counterattack traditional clustering. The method and the images of accuracy and recall rates published in this article, the specific results are shown below (Fig. 10).

For the same sample data set, four different algorithms are used to analyze it, and the sensitivity and specificity of different models are calculated according to the result distribution, and four ROC curves are obtained. The ROC curve can graphically describe the relationship between the true rate and false positive rate of a classifier, so it is often used to compare the classification accuracy of different classifiers, and the closer the curve is to the top of the ordinate, the sensitivity and specificity of the sample The higher the degree, the higher the accuracy of the classifier. As shown in the figure, the method proposed in this paper is closest to the top of the ordinate, and the local anomaly factor algorithm is closest to the diagonal, so this paper has better classification performance and higher classification accuracy (Fig. 11).

In the analysis stage of the experimental results, the conventional graphical method is also used to show the situation of the experimental results. The figure shows the experimental results of anomaly detection false detection rate experiment. The statistical

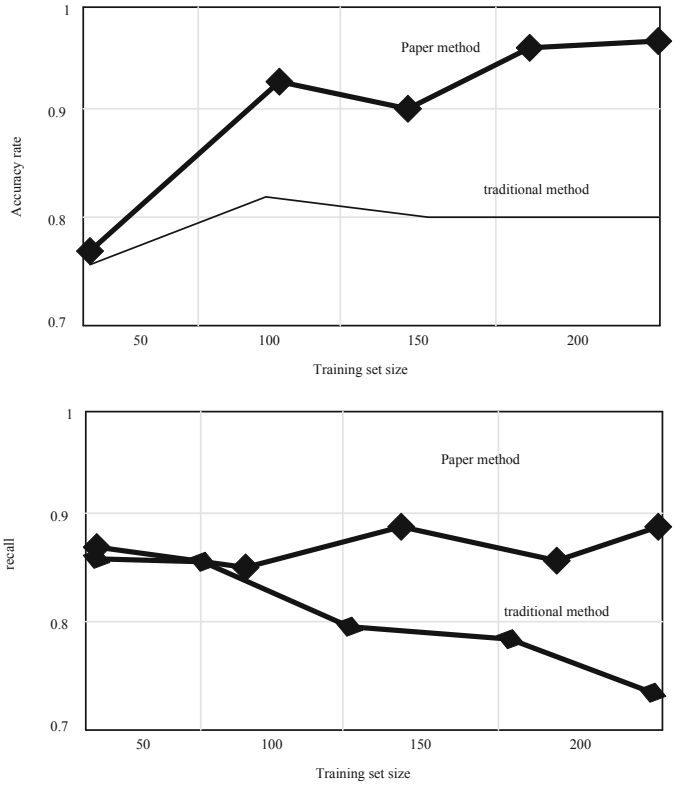

Fig. 10. Comparison test results of precision rate and recall rate

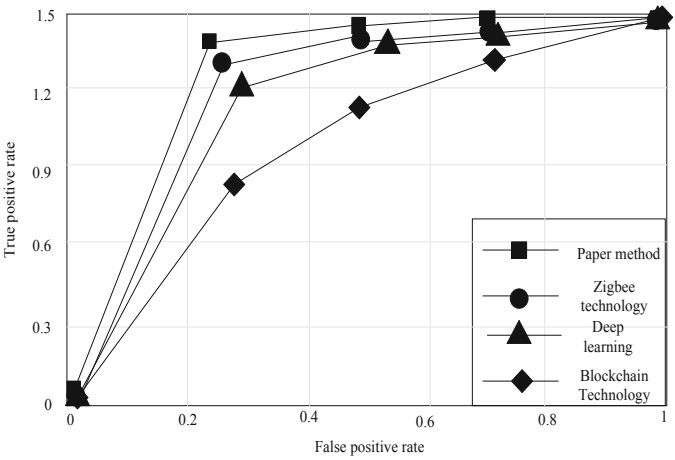

Fig. 11. ROC curve of anomaly detection algorithm

value of false detection rate is obtained by comparing experimental designs with different weights (Fig. 12).

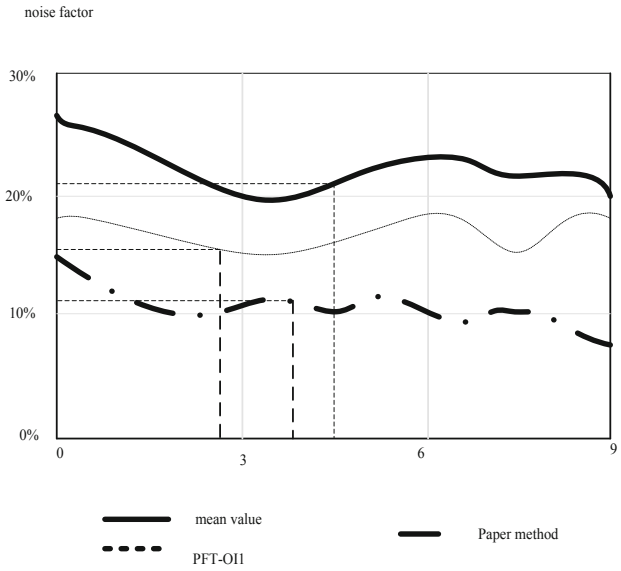

Fig. 12. Comparison test results

In the experimental stage, this method can clearly show the abnormal degree judged by the algorithm, and can more quickly and accurately achieve the research goal of accurate identification and rapid positioning of the number of teaching systems, so as to fully meet the research requirements.

4 Conclusion

Taking anomaly data detection as the research object and using cloud computing and fuzzy membership function as tools, this paper mainly makes an in-depth research on the widely used isolated forest algorithm. When the existing algorithms carry out anomaly detection, they randomly select attributes to build trees, while ignoring the anomaly degree of each data for the selected attributes. A data anomaly detection method based on cloud computing algorithm is proposed. From multiple dimensions, the membership of the detection results of each one-dimensional attribute is judged, and finally the final evaluation result is obtained by fuzzy operation with the fuzzy matrix.

This method has achieved good detection accuracy in distributed intelligent teaching system anomaly detection, but the complexity of the algorithm needs to be further simplified to effectively improve the detection time.

References

1. Lei, H.: Exploration and practice of abnormal flow detection in intelligent campus construction based on data mining. Taiwan Strait Sci. Technol. Ind. **35**(09), 55–57 (2022)
2. Zhe, C., Chong, W., Zhiqiu, H.: Program design course teaching system based on dynamic analysis. Comput. Syst. Appl. **29**(10), 114–119 (2020)
3. Yangyang, F.: Design of malicious tamper detection system for teaching system based on web crawler. Digital Commun. World **04**, 122–123 (2020)
4. Lin, P.H., Chen, S.Y.: Design and evaluation of a deep learning recommendation based augmented reality system for teaching programming and computational thinking. IEEE Access, **99**, 1–8 (2020)
5. Jun, L., Liyan, Z., Xiaoyuan, L., et al.: Cyclic redundancy check method of serial communication data flow based on ZigBee. Comput. Simul. **38**(1), 226–230 (2021)
6. Jin, R., Wei, B., Luo, Y., et al.: Blockchain-based data collection with efficient anomaly detection for estimating battery state-of-health. IEEE Sens. J. **99**, 1 (2021)
7. Wójcik, K., Piekarczyk, M.: Machine learning methodology in a system applying the adaptive strategy for teaching human motions. Sensors **20**(1), 314–326 (2020)
8. Xiaoming, L., Yi, Y., Yue, Z.: Simulation of large data multi-resolution acquisition method based on Java3D network. Comput. Simul. **37**(2), 5–18 (2020)
9. Kumar, R., Patil, O., Karthik, N.S., et al.: A machine vision-based cyber-physical production system for energy efficiency and enhanced teaching-learning using a learning factory. Procedia CIRP **98**(5), 424–429 (2021)
10. Ll, A., Yu, P.A., Lwb, C., et al.: Improving EGT sensing data anomaly detection of aircraft auxiliary power unit. Chin. J. Aeronaut. **33**(2), 448–455 (2020)
11. Wu, L.: Student model construction of intelligent teaching system based on Bayesian network. Pers. Ubiquit. Comput. **24**(3), 419–428 (2020)
12. Jiong, G., Qian, R., Jianjiang, H.: A summary of foreign research on the application of artificial intelligence in teaching. Audio Vis. Educ. Res. **41**(2), 9–18 (2020)
13. Jian, M., Ligang, N., Xin, L.: Design of laboratory intelligent teaching system in key universities based on multimedia and network technology. Mod. Electron. Technol. **44**(20), 6–16 (2021)