Design of Network Traffic Anomaly Monitoring System Based on Data Mining

Yanling Huang[1(✉)] and Liusong Huang[2]

[1] Department of Computer and Art Design, Henan Vocational College of Light Industry, Zhengzhou 450001, China
kjkmgm123@163.com

[2] Software Engineering, Maanshan Teacher's College, Ma'anshan 243041, China

Abstract. The security hidden dangers in the network will affect the normal operation of the network. Therefore, in order to better ensure the security of the network structure, it is necessary to monitor the abnormal network traffic. However, due to the low monitoring accuracy and long monitoring time of the traditional network traffic anomaly monitoring system, this paper designs a network traffic anomaly monitoring system based on data mining. Through the configuration of data acquisition equipment, analysis equipment, exception handling equipment and system management equipment, the hardware structure of the system is designed. On this basis, through the system software functions of acquisition module, data processing module, data analysis module, data application module and infrastructure management module, the abnormal monitoring of network traffic is realized through data mining. Finally, the experiment proves that the network traffic anomaly monitoring system based on data mining has higher monitoring accuracy and shorter monitoring time, which is practical in practical application and fully meets the research requirements.

Keywords: Data mining · Network flow · Abnormal monitoring

1 Introduction

With the continuous enrichment of network functions, the potential security problems in the network are becoming increasingly prominent, which puts forward higher requirements for network security monitoring [1]. Abnormal traffic is a common security risk faced by the network. Insufficient configuration of monitoring equipment or software and the new compiled virus formed by combining computer virus and hacker technology are the main reasons for the frequent occurrence of abnormal network traffic. In order to ensure the safety of network applications, it is necessary to find problems in time and quickly monitor the fault location through real-time monitoring of network traffic, so as to avoid network failure to the greatest extent. Network traffic data information is growing rapidly in the era of big data. The content dissemination and learning process of network traffic data information need to be realized by combining network

W. Fu and L. Yun (Eds.): ADHIP 2022, LNICST 468, pp. 549–563, 2023.
https://doi.org/10.1007/978-3-031-28787-9_41

technology and information technology [2]. The large-scale network based on complex network structure covers computers and network equipment of different manufacturers. Most of the different functions or applications of the network need to be realized based on different protocols [3]. In the network environment composed of different personalized learning networks, channel congestion and unbalanced distribution are easy to occur, which makes it difficult to evenly distribute trust nodes and accurately locate the problems, resulting in abnormal traffic. Real-time and accurate monitoring of network abnormal traffic has become the main means to improve network security.

The network traffic abnormal monitoring system based on Android platform is designed by traditional methods, and the key technologies of traffic monitoring are analyzed. The designed traffic monitoring system includes data extraction module, traffic monitoring module, data management module and user interface module. Finally, the system function is tested The results show that the module can meet the requirements of mobile phone traffic monitoring, but the accuracy of abnormal network traffic monitoring is low. Based on this, this paper designs a network traffic anomaly monitoring system based on data mining, optimizes the hardware configuration of the system, and combines data mining technology to optimize the software function and smooth operation of the system, improve the identification efficiency of network traffic anomalies, and improve the monitoring function of network traffic anomalies. Finally, the experiment proves that the system designed in this paper has high practicality.

2 Network Traffic Anomaly Monitoring System Based on Data Mining

2.1 Overall System Architecture

The network traffic anomaly monitoring system adopts a layered architecture design, which divides the whole system into six layers: acquisition layer, access layer, computing layer, storage layer, service layer and application layer. Each layer has its own independent functions, only depends on the services provided by the next layer to itself and provides its own services to the upper layer. The upper and lower layers coordinate and cooperate with each other to provide complete functions for the whole system. The overall architecture of the system in this paper is shown in Fig. 1.

The layered architecture meets the principle of single responsibility, allowing each layer to focus on its own functions, and the responsibility boundary is clear. It can not only simplify the system design, but also comply with the object-oriented design principle of high cohesion and low coupling. All the same types of businesses are placed on the same layer, and the upper services only rely on the functions implemented by the lower layer. At the same time, the lower layer shields its own internal implementation from the upper layer, as long as the service interface provided by the lower layer to the upper layer remains unchanged, Developers' modifications to the lower layer will not affect the upper layer services [4].

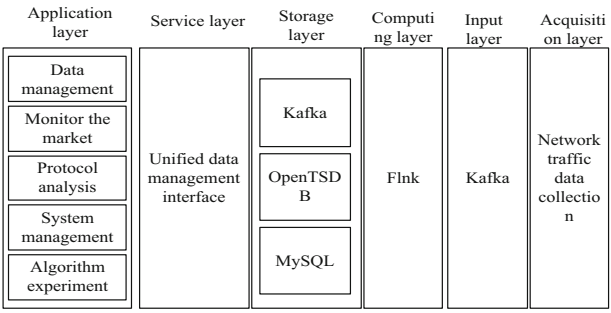

Fig. 1. Overall architecture of the system

2.2 Hardware Structure Design of Network Traffic Anomaly Monitoring System

2.2.1 Hardware Configuration Structure

The hardware structure of network traffic anomaly monitoring system is designed to better visualize the data results in the system, make the data intuitive and clear, and facilitate users to understand and use [5]. It is a complete process to filter and evaluate the results based on data mining technology, present the correct results by using visual methods, mine previously unknown, effective and practical information from large databases, and use this information to make decisions or enrich knowledge. The data mining process is shown in Fig. 2:

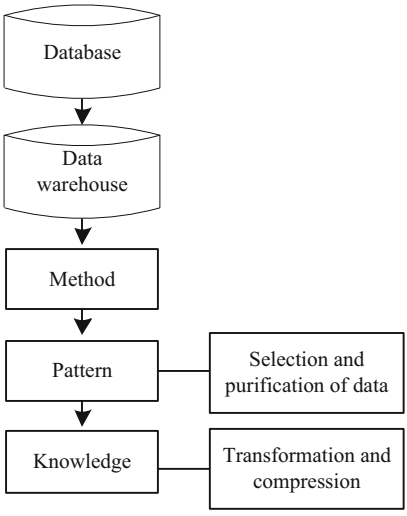

Fig. 2. Network traffic data mining process

In order to ensure the data mining effect of network traffic, it is necessary to optimize the hardware structure of the system, optimize the hardware structure of the monitoring system based on the characteristics of multi-functional network, and improve the convenience of user access. The block diagram is shown in Fig. 3:

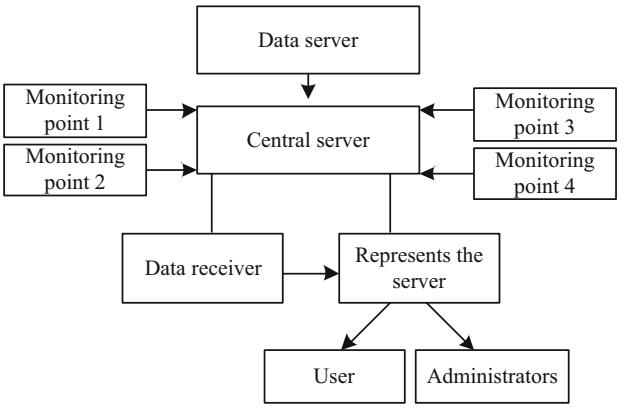

Fig. 3. Hardware configuration block diagram

The system includes four independent and interrelated equipment configurations: data acquisition equipment configuration, analysis equipment configuration, exception handling equipment configuration and system management equipment configuration [6]. The collection equipment is configured to provide the original flow data for the whole system, which is the basis of the whole system. The acquisition device configuration is responsible for collecting data from the set netf1ow output device, storing it in the database after preprocessing, and submitting it to the analysis device configuration for further analysis and processing. In the database, a source IP address table is generated according to the original data to save frequent source IP addresses. The purpose of analyzing equipment configuration is to conduct real-time and long-term monitoring of the new IP address, establish a reasonable analysis model according to the change law of the new IP address, and conduct real-time and long-term analysis and reasonable prediction of the data. The exception handling equipment configuration is responsible for the alarm of abnormal phenomena, and cooperates with the network firewall and network management system to maintain the availability of the network [7].

2.2.2 System Hardware Architecture

The overall design of the hardware architecture of the monitoring system is shown in Fig. 4.

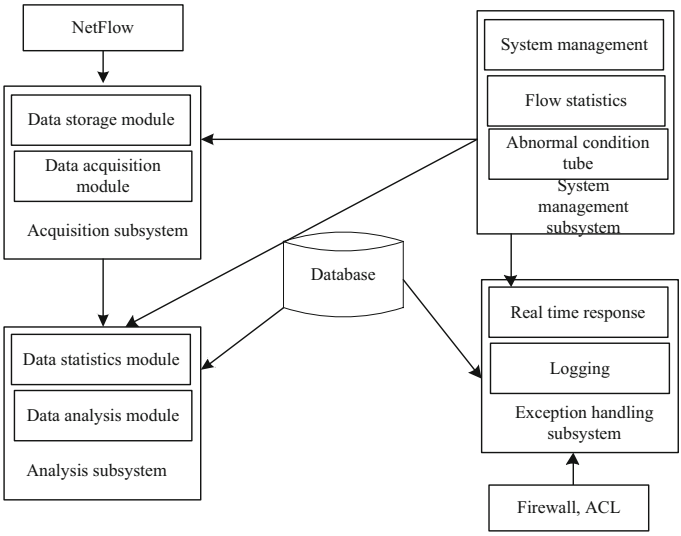

Fig. 4. Hardware architecture of monitoring system

The figure shows in detail the relationship between each equipment configuration and its role in the whole architecture, as well as the main module division of equipment configuration and the relationship between them. D realize distributed computing through RM technology of Java, and run different functional modules on different hosts, so as to make full use of network resources and computing power of multiple hosts and reduce costs [8]. DOS monitor can only be activated when monitoring abnormal network traffic. I filter will be activated when monitoring DoS attack. Therefore, in the case of positive traffic, the DOS monitor is not required. At this time, the system RMI server (data storage server) will learn the new IP address as the database middleware. Just load the JDBC actuator on the server side, and the client can call the JDBC on the server side through the RMI of Java. This three-tier structure realizes the isolation of functional modules, reduces the load of the client and server, and makes the load distribution of the whole system more reasonable. Users download the web page containing applet from the server through the browser, run swing of the base applet on the user's host, and communicate with the server process on the system configuration server through RM, so as to use the graphics community to manage the whole system [9]. The monitoring object mainly adopts the active monitoring method through four monitoring points to complete the effective monitoring process; After the monitored data is processed by the data server, a unified data sheet is obtained for subsequent query; The central server receives and summarizes the monitoring data obtained at each monitoring point, and manages the software configuration information; The traffic anomaly monitoring results are finally managed and presented through the presentation server.

2.3 System Software Function Optimization

According to the user's demand analysis, extract the business problems that the software system can help users solve, define use cases, that is, analyze the user's business problems, and plan the functional modules of the system. This step is the understanding and sublimation of users' business needs, indirectly reflects the system structure of network traffic monitoring and analysis system based on data mining, and lays a solid foundation for subsequent program development. Through the analysis of user requirements, a business use case diagram is established, as shown in Fig. 5:

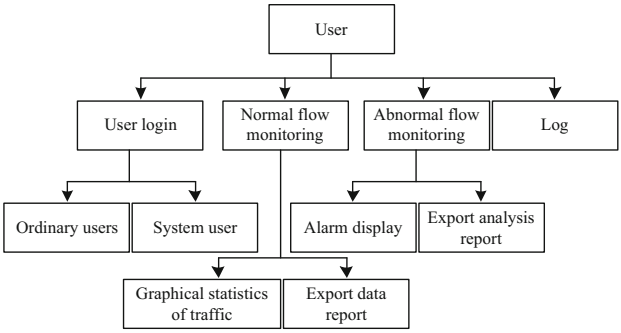

Fig. 5. Business case diagram of network traffic monitoring and analysis based on Data Mining

As shown in the above figure, users can log in, but the attributes of logged in users are divided into two categories: ordinary users and system users. Ordinary users can view the graphical statistics of data flow and obtain the alarm display of abnormal flow monitoring [10]. System users can view the detailed statistical data of individual user traffic and the analysis report of abnormal traffic. The purpose of setting system user permissions is to protect the privacy of personal PC data. The alarm of abnormal traffic and the user's operation record will be stored in the log record in real time for system users to query. Network traffic is represented as a set of time series in large-scale networks. In order to provide accurate characteristic basis for the monitoring system, the traffic characteristics can be extracted by constructing a signal model. In large-scale networks, MAC is accessed through CSMA/CA channel, and the periodic broadcast network model is used to obtain the clustering model of the corresponding data structure for the network traffic information in transmission, which is represented by $n(k)$. It is assumed that the signal received by the cognitive user is represented by $s(k)$, and the traffic information sent by the authorized user is represented by h. The expression of the channel access model is as follows:

$$x(k) = \begin{cases} n(k) & \preceq 1 \\ hs(k) + n(k), & \succ 1 \end{cases} \tag{1}$$

It is assumed that for abnormal traffic, β represents its spectrum sensing transmission signal, and $\sigma_{x_1}^2$ represents the channel gain; k represents the number of iterations; n(k) refers to additive Gaussian white noise; p refers to channel equalization index; the

variance of R during generation n is represented by i; the expression of selecting adaptive generation step is as follows:

$$\mu_1(k) = (1-\beta)/R_n^i x(k) - p\left[\sigma_{x_1}^2(k)\right] \tag{2}$$

$$\sigma_{x_1}^2(k) = \beta\sigma_{x_1}^2(k-1) + (1-\beta)x_1^2(k) \tag{3}$$

Once the network traffic is abnormal, the IP address and port distribution will change accordingly. If the network configuration is wrong, the original IP address and target IP address will increase, resulting in a sharp increase in host messages. According to this feature, the network traffic matrix method is used to analyze the dispersion of traffic distribution characteristics. Assuming that the flow characteristic is A, the total number of samples is B, and the number of occurrences of a specific flow characteristic i is n_i, therefore, the flow characteristic sample can be defined as:

$$\begin{aligned} G(x) &= -A\sum_{i=1}^{C}\left(\frac{\mu_1(k)}{B}\right) - \left(\frac{n_i}{\sigma_{x_1}^2(k)}\right) \\ B &= \sum_{i=1}^{C} n_i \end{aligned} \tag{4}$$

If the results of all selected samples are consistent, then $G(x) = 1$; If the results of all selected samples are highly dispersed. Thus, the abnormal behavior of different flow characteristics can be described, as shown in Table 1.

Table 1. Identification of abnormal flow behavior characteristics

Anomaly type	Exception definition	Abnormal characteristics
Configuration error	Device failure caused by incorrect configuration of routing port	Large abnormal characteristic value The normal eigenvalue is large
Service attack	Service attack	The abnormal characteristic value is small Normal eigenvalue is small
Burst access	Multiple hosts send traffic to a single host	Large abnormal characteristic value The normal eigenvalue is large
worm scanning	A small number of ports on the destination host are detected	The abnormal characteristic value is small The normal eigenvalue is large

The monitoring and analysis module is used to subdivide the abnormal data traffic in the network into unknown abnormal data and known abnormal data, which is conducive to improving the alarm efficiency and the work efficiency of the network administrator. In its working process, it first receives standardized abnormal data, and then uses data mining algorithm to calculate the similarity of abnormal data. Those with high similarity are classified into one class, and all kinds of direct data association or similarity are relatively low. Judge whether it is necessary to send relevant abnormal data to the alarm module according to the danger value K. if the similarity between the data packet and the same class is greater than the clustering radius K value within a certain period of time, it indicates that it has exceeded the early warning range, Then send the relevant characteristic data to the alarm module. The alarm module displays and alarms the data in detail. Figure 6 shows the operation process of network data monitoring and analysis module:

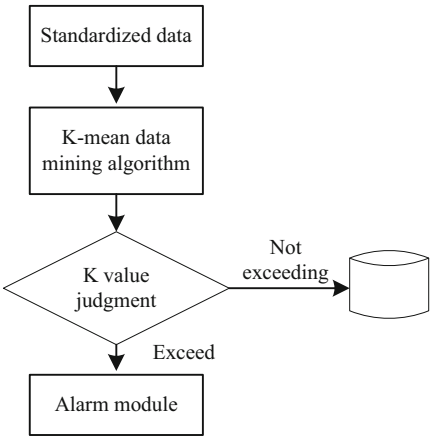

Fig. 6. Operation flow of network data monitoring and analysis module

Using the important characteristic of network traffic - chaos, the model is mainly used to analyze the self similar characteristics of network traffic, and analyze and predict from the perspective of non-linearity of traffic. Because in practice, the flow data is limited and its phase space trajectory is complex and changeable, it is difficult to be applied in reality, and it still stays at the level of theoretical research. Compared with the global method, the local method has stronger practicability and less calculation, but it still has some defects: on the one hand, it needs more storage space in calculation, on the other hand, it needs more time to solve model parameters and construct adjacent state vectors. The characteristics and applicable scenarios of some representative network traffic prediction models mentioned above are summarized in Table 2:

Table 2. Network abnormal traffic prediction model

Model	Complexity	characteristic	Applicable scenario
AINA	low	The prediction of short correlation flow is more accurate	Online prediction and wireless sensor networks
WNE	high	For non-stationary flow, the prediction is more accurate	It has high requirements for prediction accuracy and prediction time, and has sufficient resources to support the network
ABBGE	high	It is more accurate to predict the short-term and long-term related flow	
CO	low	It can better describe the change of sample trend and predict higher	
AII	high	The nonlinear flow prediction is more accurate	
SON	high	For small samples, the flow prediction is more accurate and has strong generalization ability	

The negative impact of abnormal flow on the network is self-evident. Therefore, timely and effective monitoring of abnormal flow wells and taking appropriate methods to control them have become an indispensable part of the network flow monitoring system. According to the different locations of anomalies in the network, combined with the characteristics of data mining technology, the system subdivides the module into abnormal traffic monitoring for the controller and abnormal traffic monitoring for the host.

2.4 Implementation of Abnormal Network Traffic Monitoring

Combined with the system architecture design, the system is divided into five functional modules, namely data acquisition module, data processing module, data analysis module, data application module and infrastructure management module. Each functional module can be divided into smaller functional modules, and each functional module is responsible for a single system function. Describe the design of the five functional modules of the system and their subdivided functional modules, as shown in Fig. 7.

Because the two-stage abnormal traffic monitoring model needs to be applied in the scenario of data imbalance, the abnormal traffic belongs to a few categories, and the data of abnormal traffic is much less than that of normal traffic. Therefore, based on the original two-stage abnormal traffic monitoring model, the data is oversampled by data mining algorithm to avoid the problem of model over fitting caused by unbalanced data set as far as possible. The abnormal flow monitoring model at different stages is shown in Fig. 8:

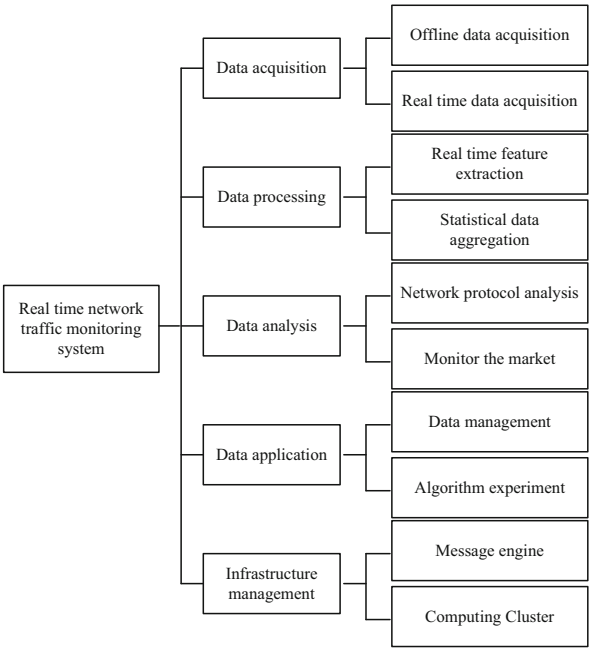

Fig. 7. System function module structure optimization

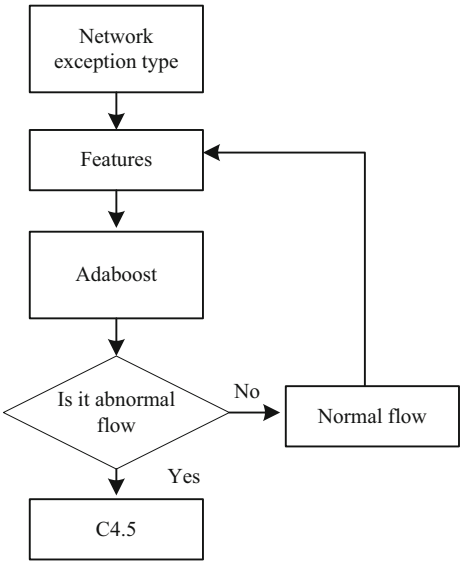

Fig. 8. Abnormal flow monitoring model in different stages

Monitoring is an important function of real-time traffic monitoring system, which is mainly responsible for the visual display of all indicators of monitoring network and key indicators of infrastructure on which the system depends. Users can simply and quickly organize a monitoring market by creating, organizing, managing data sources and configuring forms. By monitoring the market, system users can clearly and intuitively understand the operation of the whole monitoring network and system. If abnormal behavior of network traffic is found, it can be found directly in the monitoring market. For experienced network security operation and maintenance personnel, they can also find potential network security threats from various statistical charts of the monitoring market, and then update the network abnormal traffic monitoring model and network defense response module in time, Increase the monitoring ability of the system to network abnormal traffic, facilitate users to operate and troubleshoot the system, and improve the operation and maintenance efficiency of the system.

3 Analysis of Experimental Results

A series of performance tests are carried out on the system designed in this paper to ensure that the system fully meets the system requirements. The network traffic uses the same configured server in the same computer room. The specific experimental parameters are shown in Table 3.

Table 3. Experimental platform server parameters

Hardware configuration	CPU	Pentium MMX200
	Memory	64G
	Hard disk	300G
Software configuration	Operating system version	Linux 6.5
	Storm version	Storm 0.95
	Kafka version	Kafka 0.8.1
Network environment	Network card	Gigabit Ethernet

The function test mainly focuses on whether the function of the system meets the expectation, and verifies whether the logic and function of each functional module meet the system requirements. The system mainly adopts the methods of unit test and integration test to test the function of the system. The unit test mainly verifies whether the functional logic of the system is correct and whether all exception handling is reasonable. The integration test mainly verifies whether the function of the system meets the system requirements. For example, the test results of the data acquisition module are shown in the table, and the test results meet the expectations, as shown in Table 4:

Table 4. System data acquisition and transmission function test

Test name	Offline and real-time traffic collection function test
Test purpose	Test whether the offline traffic collection script and real-time traffic collection script can collect network traffic and whether the real-time script works normally
Preconditions	Offline traffic packet integrity
testing procedure	Offline traffic collection script starts locally; System registration monitoring equipment; Issue the real-time traffic collection script and start it; Kill real-time traffic collection script
Expected results	Offline traffic collection script and real-time traffic collection script work normally
Test result	The test results are in line with the expected results

In order to test the anomaly monitoring ability of the system to known attack types. In the experiment, we launched eight kinds of network attacks during training from an attack host (202.201.94.190) of network VIAN1 to the host (202.201.95.1–202.201.95.20) of network VAN2. The final monitoring results are shown in Table 5.

Table 5. Traffic monitoring results of known attack types

Data type	Normal data	Tcpport scan	Vulnerability scanning	OS scan	Tcpsyn scan	ACK scan	Land attack	Syn flood	Ddos	Total
Number of records detected	13839	3619	2918	1232	3292	0	38	5575	5749	37491
Actual records	14000	4000	3070	1430	4000	450	90	5760	6000	38800

The system performance test is carried out after the system function test. On the premise that the system function meets the expectation, the performance test is carried out to collect the indicators of the server and application and the use of hardware resources. The performance test will focus on the system throughput, end-to-end delay, system performance index and fault recovery speed of the cluster. The test results are shown in Table 6:

Table 6. System monitoring performance test results

Test object	Performance index	Test result
Kafka	Average throughput of producers	45 MB/s
	average delay	1.3 s
	Average consumer throughput	129 MB/S
	Consumer rebalancing delay	9.3 s
Flink	Average End-to-End Delay	4.5 s
system	Average response time	2.5 s
	QPS	1300
The server	CPU utilization	79%
	Memory	95%
	Disk throughput	420 MB/s
	Mean recovery time	6 min 15 s

In order to verify the effectiveness of the multi-function network monitoring system with the traditional traffic monitoring system, as shown in Fig. 9.

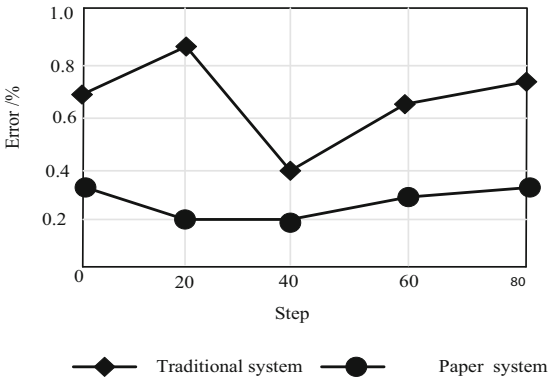

Fig. 9. Comparison results of network abnormal traffic monitoring errors of the two systems

It can be seen from this that the error of the system proposed in this paper in real-time monitoring of abnormal traffic problems in the network is within 0.2–0.4, while the error of the traditional system in monitoring abnormal traffic in the network is within 0.4–0.9. The error of the system in this paper in real-time monitoring is smaller, and the monitoring effect is better, which fully meets the research requirements.

In order to further verify the effectiveness of the method in this paper, the time taken by the system in this paper and the traditional system to monitor abnormal network traffic is compared and analyzed. The comparison results are shown in Table 7.

Table 7. Monitoring time of abnormal network traffic/s

Number of experiments/time	Article system	Traditional system
10	2.3	11.5
20	2.6	13.4
30	3.4	14.7
40	3.9	14.9
50	4.2	15.2
60	4.9	15.9
70	5.6	16.8

According to the data in Table 7, the time taken by the system in this paper to monitor the abnormal network traffic is within 5.6 s, while the time taken by the traditional system to monitor the abnormal network traffic is within 16.8 s. The time taken by the system in this paper to monitor the abnormal network traffic is the shortest and the monitoring efficiency is the highest.

4 Concluding Remarks

Most of the traditional network abnormal traffic monitoring systems are vulnerable to noise interference and large error of monitoring results. Therefore, this paper designs an optimization scheme of network abnormal traffic monitoring system based on data mining to realize the real-time and effective monitoring process of abnormal traffic. The test results show that the optimization scheme in this paper has high monitoring accuracy and significantly reduces the abnormal traffic monitoring error. However, the operating efficiency of the system has not reached the expected effect. Therefore, in the next research, the algorithm will be further improved to shorten the operation time and improve the operating efficiency of the system.

References

1. Ifzarne, S., Tabbaa, H., Hafidi, I., et al.: Anomaly detection using machine learning techniques in wireless sensor networks. J. Phys: Conf. Ser. **1743**(1), 012021–012034 (2021)
2. Yang, J., Hou, X.: Research on data mining algorithm based on positive and negative association rules. Comput. Technol. Dev. **30**(11), 64–68 (2020)
3. Choi, H., Kim, M., Lee, G., et al.: Unsupervised learning approach for network intrusion detection system using autoencoders. J. Supercomput. **75**(9), 5597–5621 (2019)
4. Lian, J., Fang, S., Zhou, Y.: Model predictive control of the fuel cell cathode system based on state quantity estimation. Comput. Simul. **37**(07), 119–122 (2020)
5. Xu, Y., Sun, Z.: Research development of abnormal traffic detection in software defined networking. J. Softw. **31**(01), 183–207 (2020)

6. Xiao, F., Chen, L., Zhu, H., et al.: Anomaly-tolerant network traffic estimation via noise-immune temporal matrix completion model. IEEE J. Sel. Areas Commun. **37**, 1192–1204 (2019)
7. Ma, W., Zhang, Y., Guo, J.: Abnormal traffic detection method based on LSTM and improved residual neural network optimization. J. Commun. **42**(05), 23–40 (2021)
8. Meng, Y., Qin, T., Zhao, L., et al.: Network anomaly detection method based on residual analysis. J. Xi'an Jiaotong Univ. **54**(01), 42–48+84 (2020)
9. Peng, Y., Chen, X., Chen, S., et al.: Cross-domain abnormal traffic detection based on transfer learning. J. Beijing Univ. Posts Telecommun. **44**(02), 33–39 (2021)
10. Liu, Y., Li, J., Zhang, Y., et al.: Network abnormal flow detection method based on feature attribute information entropy. Netinfo Secur. **21**(02), 78–86 (2021)