Abnormal Traffic Detection Method of Educational Network Based on Cluster Analysis Technology

Lei Ma(✉), Jianxing Yang, and Fayue Zheng

Beijing Polytechnic, Beijing 100016, China
malei235@tom.com

Abstract. Due to the long detection time of the existing education network abnormal traffic detection methods, the detection accuracy of individual abnormal traffic information is relatively low, which is easy to threaten the operation security of the education network. Therefore, an education network abnormal traffic detection method based on cluster analysis technology is proposed. According to the standardization principle, the key abnormal traffic information is processed, and then according to the definition of subspace clustering, the specific numerical results of the cluster similarity index are calculated to complete the clustering and analysis of the abnormal traffic of the education network. On this basis, execute the abnormal flow information extraction instruction, combine the known median absolute deviation measurement conditions, analyze the minimum covariance results of the detection results, and realize the smooth application of the education network abnormal flow detection method based on the cluster analysis technology. The experimental results show that, compared with traditional detection methods, under the effect of cluster analysis technology, the maximum value of abnormal traffic information of education network can reach 14.1×10^{-7} T per unit time, which is in line with the reality of rapid detection of abnormal traffic information of education network Application requirements can better avoid the threat and impact of abnormal information parameters on the security of the education network.

Keywords: Cluster analysis · Network flow · Anomaly detection · Standardization of specifications · Median deviation · Minimum covariance

1 Introduction

Flow detection refers to the monitoring of data flow, usually including the speed of outgoing data, incoming data and total flow. Traffic detection sometimes refers to monitoring and filtering the user's data traffic to effectively master the bad information within the monitoring range, which is commonly used in the professional term of network security. From the perspective of the whole, application and other aspects of network traffic. Through the analysis of data and indicators reflecting the real situation of the network,

W. Fu and G. Sun (Eds.): eLEOT 2022, LNICST 453, pp. 724–736, 2022.
https://doi.org/10.1007/978-3-031-21161-4_55

such as exit bandwidth utilization, application traffic ranking, application traffic ranking, main traffic flow direction, main use ports, network quality, connection success rate and data retransmission rate, it provides a basis for setting traffic management strategy [1]. All real-time analysis is refreshed every fixed time to facilitate the timely discovery of network problems; It can quickly locate the host or application with abnormality or failure through several mouse clicks. Multiple management devices can be virtualized on one device, each virtual device manages a dedicated line, and one device can realize independent analysis, guarantee and management of multiple dedicated lines at the same time.

Cluster analysis refers to the analysis process of grouping a collection of physical or abstract objects into multiple classes composed of similar objects. It is an important human behavior. The goal of cluster analysis is to collect data to classify on the basis of similarity. Clustering comes from many fields, including mathematics, computer science, statistics, biology and economics. In different application fields, many clustering technologies have been developed. These technologies and methods are used to describe data, measure the similarity between different data sources, and classify data sources into different clusters. Clustering is a process of classifying data into different classes or clusters, so the objects in the same cluster are very similar, while the objects in different clusters are very different [2]. For the abnormal traffic detection and screening behavior in the education network, although the traditional application methods can define the location of data information, the mutual measurement relationship between information and information is not emphasized. The detection time for individual abnormal traffic information is long and the accuracy level is relatively low, which makes the education Internet vulnerable to the threat of abnormal traffic information during the operation. In order to solve the above problems, this paper proposes a new abnormal traffic detection method based on cluster analysis technology.

2 Cluster Processing and Analysis of Abnormal Traffic in Education Network

Starting from three perspectives: standardization of abnormal traffic information specifications, subspace clustering processing, and cluster similarity calculation, the clustering processing and analysis of abnormal traffic in the education network are completed.

2.1 Standardization of Abnormal Flow Information

For random data objects in the education network, the corresponding traffic data may have multiple attributes at the same time, and the dimension levels between these attribute values are not completely different, which indirectly leads to the quantitative indicators of different attributes. The difference in numerical values is very large, which in turn causes those specifications with larger orders of magnitude to have a great impact on the clustering results, while specifications with smaller orders of magnitude have little effect on the clustering results, which is also caused by clustering errors. The direct cause [3]. In order to solve the problem caused by the non-uniform dimension, the abnormal traffic information of the education network to be detected should be dimensionlessly

operated before clustering, so that the value of each attribute is kept within a uniform numerical range. It is stipulated that the value of abnormal traffic information s of the education network always belongs to the set of natural numbers N, P_s represents the clustering execution intensity index of the information s, v_0 represents the information attribute discriminant value under the action of the clustering algorithm, and v_s represents the attribute discriminant value of the information s. Combining the above physical quantities, the standard processing principle of the clustering specification of abnormal traffic information of the education network can be expressed as:

$$J_s = \frac{\sum_{s=1}^{+\infty} P_s \sum_{s=1}^{+\infty} ||v_s - v_0||^2}{r_s - r_0} \tag{1}$$

In the above formula, r_0 represents the random variable definition coefficient of the abnormal flow information of the education network under the effect of clustering, and r_s represents the variable definition coefficient of the abnormal flow information s of the education network under the effect of clustering.

Cluster analysis is neither a formula nor a random application idea, but a complete data information processing process. Input the data into the established processing environment, and the expected results will be obtained after the operation of clustering law. Clustering processing is a complex process with many steps, and each step needs to be repeated constantly to achieve accurate results. Because the clustering analysis algorithm only has the ability of hard division, each data can only belong to one cluster category in the operation process. To measure whether the clustering processing method can distinguish the abnormal traffic information of education network, we should not only standardize, but also retain the initial resolution of the original attributes.

2.2 Subspace Clustering

Subspace clustering is also called the selection of information parameter feature space. For the education network environment, the initial abnormal traffic information storage space is divided into many different subspaces. On this basis, only the important subspaces need to be clustered.

The clustering analysis algorithm usually uses greedy value to define different feature subspaces, then evaluates these subspaces storing abnormal traffic information of education network through corresponding measurement criteria, and finally obtains the required clustering conditions. To sum up, the idea of these subspace processing is to find those sparse regions in the feature subspace, then remove the sparse regions, and the remaining data is dense regions, which is the desired clusters [4]. In other words, clustering is only carried out in the feature subspace, and those unimportant spaces are artificially ignored. Therefore, the final operation result of clustering algorithm may not be complete, but the ignored result can only reflect the existence form of a small part of abnormal traffic information in education network.

The subspace is carried out on the basis of considering the impact of each dimension on the clustering results. It not only does not lose the impact of any dimension on the clustering results, but also ensures the real-time storage capacity of abnormal traffic

information in the education network. At present, the clustering processing of some subspaces is to map the data to each one-dimensional space for clustering, and then merge the clustering results of each space according to some law to obtain the final clustering results. However, due to the clustering of each space, the amount of calculation is ten times large, especially for high-dimensional data, and when the data is reduced to one-dimensional, The authenticity of abnormal traffic information in some educational networks may be affected. Therefore, in order to make the final detection results have strong reference value, the subspace parameters of data information should be clustered in a high-dimensional environment [5].

Let a denote the dimensionality definition coefficient of the clustering subspace, $\dot{e}$ denote the normal vector of the abnormal flow information of the education network to be checked, and e' denote the inverse function of the normal vector $\dot{e}$. Combining the above physical quantities, the subspace definition condition can be expressed as:

$$D = \sum_{a=1}^{+\infty} (\dot{e})^{|a-1|^2} / \sum_{a=1}^{+\infty} (e')^{\sqrt{a^2+1}} \tag{2}$$

It is stipulated that c_1 and c_2 represent two unrelated educational network abnormal flow information definition items, and the improper condition of $c_1 \neq c_2$ is always established, d represents the dense planning coefficient of the data information parameter in the subspace environment, and α represents the data information in the subspace environment The sparse planning coefficient of the parameter. With the support of the above physical quantities, the simultaneous formula (1) and formula (2) can express the result of subspace clustering as:

$$\tilde{D} = [J_s]^2 \cdot ||c_1 - c_2||^2 / D^{\left(\frac{\alpha}{d-1}-1\right)^2} \tag{3}$$

The cluster analysis algorithm regards a certain point in the subspace as the center. Under the condition that the unit storage amount of abnormal traffic information in the education network does not change, the parameter value less than the data information storage density at the point can be defaulted as one kind of cluster element, while the parameter value greater than the data information storage density at the point can be defaulted as another kind of cluster element, The two elements are never equal, but they can be transformed into each other through the established clustering function.

2.3 Cluster Similarity Calculation

The physical definition based on cluster similarity is as follows: given a data set with large enough sample space, it is specified that each abnormal traffic information of education network can maintain an independent corresponding relationship with a given similarity marker coefficient. On this basis, a clustering tree of data objects is constructed, and then hierarchical decomposition is carried out until certain conditions are met [6]. According to whether the hierarchical decomposition is formed from bottom-up or top-down, the similarity index can be further divided into condensed and split hierarchical clustering. Hierarchical clustering is a bottom-up strategy. Firstly, each object is regarded as a

cluster, and then these atomic clusters are merged into larger and larger clusters until all objects are in a cluster or meet a termination condition; Split hierarchical clustering is a top-down strategy. Contrary to condensed hierarchical clustering, first, all objects are placed in a cluster, and then gradually subdivided into smaller and smaller clusters until each object forms a cluster or meets a termination condition. Its specific definition form is as follows.

(1) Cohesion similarity

Firstly, by analyzing the abnormal traffic information of education network, a complete clustering condition is established, which is regarded as the basic clustering principle of information data; Then, the cohesion index of abnormal traffic information in educational network at hierarchical nodes is analyzed; Finally, the aggregation similarity coefficient is calculated.

Let $x_1, x_2, \cdots, x_n$ denote the scale values of n different abnormal flow information nodes of the education network, f_1 denotes the cohesion coefficient of the information index, β_1 denotes the cohesion index of the information parameter, and λ_1 denotes the cohesion processing authority of the information parameter. With the support of the above physical quantities, the simultaneous formula (3) can define the clustering analysis expression of agglomerated similarity as:

$$K_1 = \frac{\left|1 - \sqrt{\tilde{D} \cdot f_1 (x_1 + x_2 + \cdots + x_n)}\right|}{\beta_1^2 + \lambda_1^2} \tag{4}$$

(2) Split similarity

Firstly, the ability to split and change the abnormal traffic information of education network in the clustering environment is determined; Secondly, the matching relationship between information variables and known clustering behavior is studied; Finally, the influence of split similarity on the detection results of abnormal traffic information in education network is analyzed.

Suppose f_2 represents the splitting coefficient of the information index, β_2 represents the splitting index of the information parameter, and λ_2 represents the splitting processing authority of the information parameter. When the scale value of the abnormal traffic information node of the education network is always $x_1, x_2, \cdots, x_n$, use The above physical quantity can be defined as the expression of cluster analysis of split similarity as:

$$K_2 = \frac{\beta_2 \cdot \lambda_2}{\left|f_2 \tilde{D} / (x_1 + x_2 + \cdots + x_n)^2\right|} \tag{5}$$

In short, the basic application idea of cluster analysis technology is to divide data objects into relatively small clustering indexes through an information hierarchical way; Then, a condensed hierarchical clustering algorithm is used to find the real result cluster by repeatedly merging subclasses. In this process, not only the interconnection, but also

the approximation between clusters, especially the internal characteristics of clusters, are considered to determine the most similar sub clusters, so that it does not depend on the model provided by a static node, It can automatically adapt to the internal characteristics of the merged clusters; Finally, the global expression ability of abnormal traffic information in education network is summarized, so that the Internet host can accurately detect these information parameters [7].

3 Abnormal Traffic Detection in Education Network

With the support of cluster analysis technology, according to the processing flow of abnormal traffic information extraction, median absolute deviation calculation and minimum covariance analysis, the design and application of a new education network abnormal traffic detection method are completed.

3.1 Abnormal Flow Information Extraction

In the process of detecting abnormal traffic in the education network, due to the excessively large data base, there will be phenomena such as complex data processing, slow detection process, and inaccurate detection results. In order to avoid the occurrence of the above situation, the abnormal flow information should be accurately extracted.

First, collect the data generated by the education network attack to obtain the abnormal traffic data set. The abnormal traffic data generated is then stored in the database. Design a data pipeline for processing flow data, which is used to classify network flow data and analyze the classified flow to search for flow data with common characteristics [8]. Next, extract the data parameters that meet the cluster analysis criteria, and design

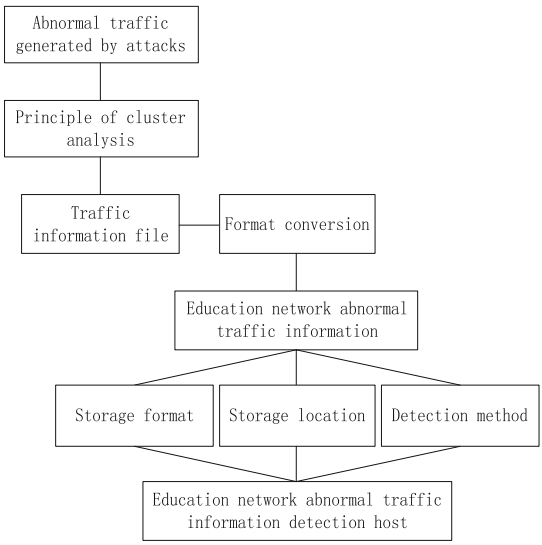

Fig. 1. The principle of the extraction of abnormal traffic information in the education network

the interface to establish a connection with the database. After receiving the data, perform abnormal flow feature selection on the detection pipeline, design a network flow data framework to process the flow, and then perform flow classification operations on the flow. The detection host performs classification operations based on the metrics set by the user, and the classification operation detects abnormal traffic in the network data. Subsequently, the classified traffic data is subjected to a stream analysis operation to analyze the attribute combination of abnormal traffic and normal traffic that is most likely to become abnormal traffic, and analyze and explain the characteristics of the traffic data. Finally, statistically analyze the data generated by the operation, provide an analysis list sorted by abnormal support, and generate a static report to display to the user. The specific extraction principle is shown in Fig. 1.

Through the cluster analysis method, the abnormal traffic information of the education network can try to create a denial of service in the detection host or service object, and the traffic is much lower than other DDoS attacks. Volumetric attacks designed to saturate the network infrastructure around the target do not only require SYN attacks that are larger than the backlog available in the target operating system [9]. If the attacker information can determine the size of the backlog and the length of time each connection remains open before timeout, the attacker can locate the exact parameters required to disable the detection host, thereby reducing the total traffic to the minimum required to create a denial of service This is also the main reason why the abnormal traffic data of the education network can be accurately extracted.

3.2 Median Absolute Deviation

For the individual abnormal traffic data of education network, the variation of clustering analysis algorithm is to use the median and median absolute deviation instead of the mean and standard deviation as the measurement of the position and scattering of detection instruction distribution. The median absolute deviation method counts the median of the absolute distance from each point in the data information sample to the median of the sample [10]. Since the median itself is resistant to outliers, each peripheral data point has a limited impact on the median absolute deviation scores of all other points in the sample.

From a statistical point of view, the median absolute deviation of abnormal traffic data in education networks is a powerful measure of the distribution of univariate quantitative data samples. If the test data is normal, the standard deviation is usually the best choice for statistical deviation. However, if the test data is abnormal, the median absolute deviation statistics can be used. For univariate data set M, $m_1, m_2, \cdots$, and m_n represent n unequal abnormal traffic information of education networks. The larger the subscript value, the greater the value of abnormal traffic information at the current location., The median absolute deviation is defined as when the absolute deviation of the median of the data is equal, that is, the value can represent the median data of the abnormal traffic information of the education network.

The specific median absolute deviation definition expression is as follows:

$$U_M = |K_1 - K_2| \frac{(m_1 + m_2 + \cdots + m_n)^2 - \xi \overline{m}}{n \times \frac{\sqrt{\phi \hat{k}}}{\hat{g}^2}} \tag{6}$$

Among them, $\overline{m}$ represents the average value of abnormal traffic information of n education networks, ξ represents the median check index, ϕ represents the predetermined deviation coefficient, $\hat{k}$ represents the most reasonable deviation measurement result of abnormal education network information, and $\dot{g}$ represents the median feature definition Permissions.

Starting from the median deviation of abnormal traffic data in education network, the median absolute deviation is the median of its absolute value. Under the function of cluster analysis algorithm, the formula is the change of average absolute deviation formula. It is less affected by outliers because outliers have less impact on the median than on the mean. In practical application, the absolute deviation of digits refers to the statistics calculated from samples. However, it can be used to estimate population parameters. Median absolute deviation is a measure of statistical deviation. In addition, the median absolute deviation can more stably count the abnormal traffic data of education network, and the statistics of abnormal values in the data set is more flexible than the standard deviation. In the standard deviation, the distance from the mean value is square, so it has a large deviation, and the abnormal value will seriously affect it. In the median absolute deviation, the deviation of a few outliers is irrelevant.

3.3 The Minimum Covariance of the Test Results

After the detection host identifies the abnormal traffic information of the education network, it needs to analyze and mine the characteristics of the data, summarize the characteristics of the abnormal data, and the frequent itemset mining method can extract the data characteristics. The combination of items that always appear together can be extracted by cluster analysis algorithm. The combination of these items is called frequent itemset. After the frequent itemset is extracted, if there are entries in a transaction, other entries can be recommended to the detection host, that is, to obtain accurate minimum covariance calculation results.

The detection of abnormal traffic in educational network refers to mining the network data information that does not conform to the normal behavior in the network flow data through the processing methods of machine learning, statistical analysis and cluster analysis. By combining the main idea of network anomaly traffic detection with the idea of big data detection and monitoring, a network anomaly detection platform based on big data can be constructed. Specifically, the data collection tasks such as log information and network traffic data information in various abnormal behaviors of educational network are completed in the data collection module; Storing heterogeneous data, and processing the burst of data through cache is realized by the storage management module; Feature extraction, statistical analysis, model training, representation extraction and full-text retrieval are completed in the intrusion behavior discovery module; Real time display, data interaction visualization, report management and system operation and maintenance management are realized in the configuration management and display module [11]. The network anomaly detection platform based on big data is shown in Fig. 2.

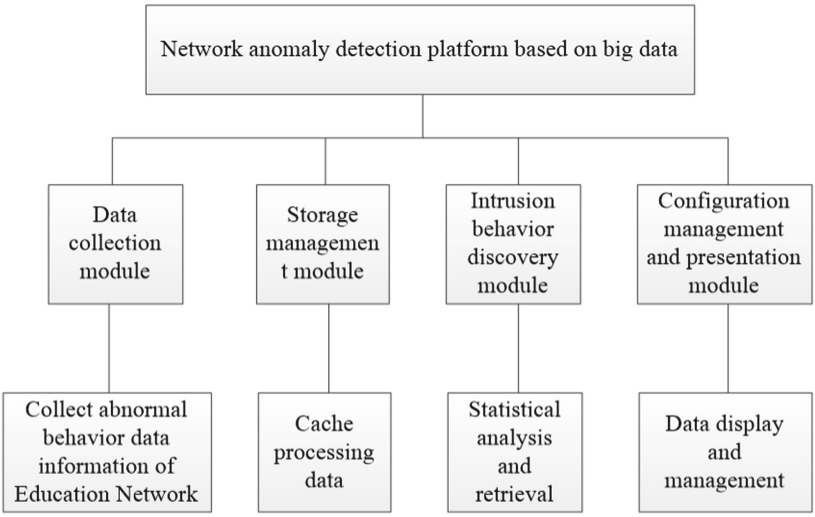

Fig. 2. Network anomaly detection platform based on big data

Let ξ represent the covariance statistical scale value of abnormal traffic information of the education network, v_0 represent the initial assignment result of the detection data indication, $\tilde{\chi}$ represent the characteristic value index of the abnormal traffic information of the education network that can be distinguished by the detection host, l_1 and h_1 represent two Different data information detection scalar. With the support of the above physical quantities, the simultaneous formula (6) can define the minimum covariance expression of the detection result of abnormal traffic information of the education network as:

$$\delta_{\min} = \left| \frac{\xi \cdot |U_M|}{\sqrt{\int_{v_0}^{+\infty} \tilde{\chi} \cdot ||l_1 + h_1|| d\tilde{\chi}}} \right|^2 \tag{7}$$

So far, the calculation and processing of various index parameters are completed, and the smooth application of abnormal traffic detection method in education network is realized with the support of cluster analysis technology.

4 Case Analysis

In order to highlight the practical application value of the abnormal traffic detection method of education network based on cluster analysis technology, the following comparative experiments are designed. Two Internet hosts with identical configurations were selected as the experimental objects. The hosts in the experimental group were equipped with the detection method of abnormal traffic in the education network based on cluster

analysis technology, and the hosts in the control group were equipped with the traditional detection method. The hosts of the experimental group and the control group were placed in the same Internet environment; The same amount of abnormal traffic data of the education network will be obtained and input into the host components of the experimental group and the control group respectively; Record the real-time detection speed of the host in the experimental group and the control group for the abnormal traffic information of the education network, and compare the experimental results with the ideal values.

The detection speed of the host component for the abnormal traffic information of the education network can reflect the degree of threat of the abnormal information parameter to the operation security of the education network. Generally speaking, the faster the detection speed, the ability to reflect the threat of the abnormal information parameter to the operation security of the education network The weaker, the stronger otherwise.

The following table records the ideal numerical situation of the detection speed of abnormal traffic information of the education network.

Table 1. Ideal value of detection speed

Time/(min)	Detection speed/($\times 10^{-7}$ T)
5	9.8
10	10.1
15	10.5
20	10.7
25	10.9
30	11.4
35	11.6
40	11.5
45	11.3
50	11.0
55	10.6
60	10.8
65	11.0
70	10.8
75	10.6
80	11.5

Analyzing Table 1 shows that with the extension of the experiment time, the ideal detection speed of the abnormal flow information of the education network shows a numerical trend that first increases and then decreases. When the time value is 35 min, the detection speed reaches the maximum value of 11.6×10^{-7} T; when the time value

is 5 min, the detection speed reaches the minimum value of 9.8×10^{-7} T, the physical difference between the two The numerical difference is 1.8×10^{-7} T.

The following figure reflects the experimental numerical changes of the experimental group and the control group's educational network abnormal traffic information detection speed (Fig. 3).

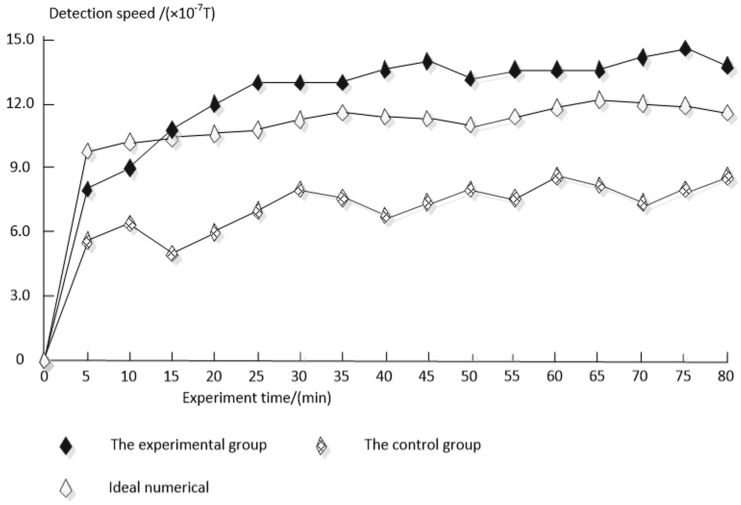

Fig. 3. Experimental value of detection speed

Experimental group: For the Internet host of the experimental group, the detection speed of abnormal traffic information of the education network shows a numerical trend of increasing, stable, rising, and decreasing. The fluctuation state in the later stage of the experiment is more obvious than that in the early stage of the experiment. When the time value reaches 15 min, the abnormal flow information detection speed of the experimental group education network is always lower than the ideal value. When the time value is equal to 45 min, the abnormal flow information detection speed of the experimental group education network reaches its maximum value, which is 14.1×10^{-7} T. Compared with the ideal maximum value of 11.6×10^{-7} T, this is an increase of 2.5×10^{-7} T.

Control group: For the Internet hosts in the control group, the detection speed of abnormal education network traffic information shows a fluctuating numerical trend. During the entire experiment, when the time value is equal to 30 min, the detection of abnormal education network traffic information The speed reached the maximum value of 7.9×10^{-7} T, which was 3.7×10^{-7} T lower than the ideal maximum value of 11.6 $\times 10^{-7}$ T, and the overall average level was much lower than the experimental group.

In order to test the superiority of this method, 500 abnormal traffic information are set in the education network, and the abnormal traffic detection method based on clustering analysis and the traditional network abnormal traffic detection method are used to detect them respectively. Taking the detection time as the experimental index, the shorter the

detection time is, the higher the detection efficiency of the method is proved. The specific results are shown in Fig. 4.

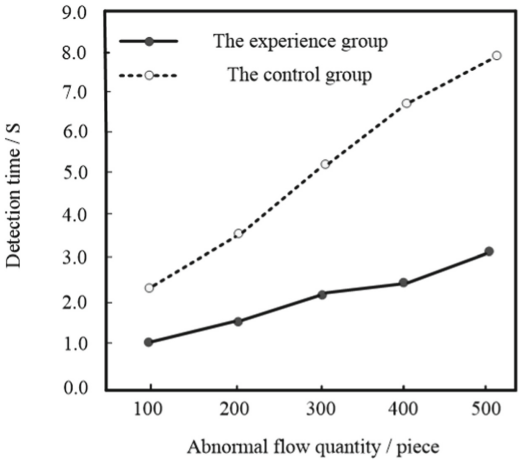

Fig. 4. Test time comparison results

It can be seen from Fig. 4 that the detection time of the two groups of methods increases with the increase of abnormal traffic data. The experimental group spent 3.0 s on detecting 500 abnormal flow data, while the control group spent 8.0 s on detecting 500 abnormal flow data. The time of the detection method proposed in this paper is shorter, which shows that the detection efficiency of this method is higher.

In summary, with the detection method based on cluster analysis technology, the Internet host's resolution speed of abnormal traffic information of the education network has been appropriately promoted, which can better solve the operational safety problem of the education network caused by abnormal information parameters. Actual application requirements.

5 Conclusion

Compared with the traditional detection methods, the detection method based on cluster analysis technology carries out cluster analysis on the subspace by standardizing the abnormal traffic information, and combines the median absolute deviation index to realize the accurate calculation of the minimum covariance of the detection results. From the practical point of view, the abnormal traffic information of education network has been detected quickly, which can avoid the threat of abnormal information parameters to the operation security of education network.

References

1. Zhao, J., Yang, Y., Xin, Y., Zhu, H.: Unsupervised network anomaly flow detection algorithm based on CGAN-LSTM. Softw. Guide **21**(03), 170–175 (2022)

2. Zhan, P., Chen, L., Cao, L., Li, X.: Network abnormal traffic detection algorithm based on characteristic symbol representation. J. Zhejiang Univ. (Eng. Ed.) **54**(07), 1281–1288 (2020)
3. Gao, M.: Network data flow anomaly detection algorithm based on mathematical model. Changjiang Inf. Commun. **34**(11), 42–44 (2021)
4. Sagatov, E., Lovtsov, K., Sukhov, A.: Identifying anomalous geographical routing based on the network delay. Int. J. Netw. Secur. **21**(5), 760–767 (2019)
5. Liu, Y., Wang, Y., Qiang, Y., et al.: Network traffic anomaly detection based on random projection and clustering. Comput. Simul. **36**(3), 289–293 (2019)
6. Passas, V., Miliotis, V., Makris, N., et al.: Pricing based distributed traffic allocation for 5G heterogeneous networks. IEEE Trans. Veh. Technol. **69**(10), 1 (2020)
7. Liu, T., Abouzeid, A.A., Julius, A.A.: Traffic flow control in vehicular multi-hop networks with data caching and infrastructure support. IEEE/ACM Trans. Netw. **28**(1), 1–11 (2020)
8. Shridhar, V.S.: The India of Things: Tata Communications' countrywide IoT network aims to improve traffic, manufacturing, and health care. IEEE Spectr. **56**(2), 42–47 (2019)
9. Mesquita, L.A.J., Assis, K.D.R., Almeida, R.C.: Multi-period traffic on elastic optical networks planning: alleviating the capacity crunch. J. Supercomput. **77**(6), 5468–5491 (2020). https://doi.org/10.1007/s11227-020-03493-7
10. Bianchin, G., Pasqualetti, F.: Gramian-based optimization for the analysis and control of traffic networks. IEEE Trans. Intell. Transp. Syst. **21**(7), 3013–3024 (2020)
11. Al-Najjar, A., Khan, F.H., Portmann, M.: Network traffic control for multi-homed end-hosts via SDN. IET Commun. **14**(19), 3312–3323 (2020)