Journal of Science and Engineering Research
Vol. 1, No. 1, 2022, pp. 63-78
eISSN: 2957-7187

Detection of Abnormal Behavior of Network Users

K. S. Geethu[1] and A. V. Babu[2]

[1,2]Department of Electronics and Communication Engineering, National Institute of Technology Calicut, Kerala, Calicut, India
[1]geethu_p120081ec@nitc.ac.in

DOI: http://dx.doi.org/10.56828/jser.2022.1.1.1
Article history: *Received (September 10, 2021); Review Result (October 20, 2021); Accepted (December 15, 2021)*

Abstract: Users' access to information has increased quickly along with the quick development of technologies like big data and the Internet of Things. However, while enjoying huge technological dividends, all sectors of society are also faced with problems brought about by information security. Anomaly detection and analysis of user access log data are one of the research hotspots in academic circles. However, the traditional anomaly detection methods for large-scale distributed data have some shortcomings. First, the collected web log data sets are time-effective, but little attention is paid to them in traditional anomaly detection algorithms. Secondly, training existing historical normal users to access data requires a lot of costs, and the anomaly detection efficiency is low. Thirdly, in the era of big data, network traffic is characterized by large data volume, high dimensions of characteristic attributes, and a large correlation between attributes. Using traditional anomaly detection methods, there will be problems of low detection efficiency and long detection time. Therefore, how to quickly and efficiently detect the anomaly of large-scale network user behavior data collected by big data platforms has become a huge challenge. Aiming at the above problems, this paper analyzes the advantages and disadvantages of various anomaly detection methods and puts forward two anomaly detection methods based on data mining, which realize high-efficiency anomaly detection. Aiming at the characteristics of a large amount of network traffic data and high data dimension, this paper proposes an improved GRU anomaly detection method. Firstly, principal component analysis is used to reduce the dimension of large-scale network traffic data sets and extract effective attributes. Then, the processed training data set is used to train the GRU-SVDD classifier model. Finally, the actual traffic to be detected is input into the GRU-SVDD comparator, and the anomaly in the traffic is detected. Aiming at the data set of network user behavior collected on the big data platform, a multi-layer protection model from the application layer and network layer is constructed. It can effectively protect the security of the big data platform, and the corresponding algorithm is designed and implemented in this project system.

Keywords: Anomaly detection, Frequent pattern mining, GRU, Principal component analysis

1. Introduction

Technologies like big data and the Internet of Things have developed quickly as a result of the big data age. All types of enterprise enterprises now operate primarily in distributed server architecture. All kinds of users' access and service provision make the system application

more and more reliable, and the analysis of user log or system service log information becomes more and more important. The log information in the system is distributed. The conventional approach is to consolidate all of the dispersed environment's log information onto a single computer for analysis and processing. I then employ frequent sequence patterns to collect all of the state information required for system operation and maintenance. At the same time, for the sake of user privacy protection and data security, all kinds of log data in a large-scale distributed environment should not be stored in different places, nor should they occupy network bandwidth and be transmitted to the same place for centralized processing. With time, the data grows faster and faster, and there are more and more data. The traditional methods are no longer suitable for the explosive growth of data. The increasing data will also increase the probability of platform anomalies, which will lead to frequent network accidents. According to the statistics of 2018, the number of Internet users in China reached about 772 million, of which the number of Internet users using mobile phones reached about 755 million. The Internet has had a great impact on our personal life. With the gradual improvement of the underlying network defense system [1], the basic communication network security protection has also been improved, but at the same time, the network security threats faced by enterprises and the active Internet of ordinary users have become more terrible. For example, in recent years, network hackers launched distributed denial of service (DDoS) attacks by exploiting the defects of Internet transmission protocols, which paralyzed distributed servers in some enterprises on a large scale to profit from them; During holidays or peak tourist season, a large number of users visit the server website, resulting in large-scale access flow, which may lead to the collapse of distributed servers and the loss of server data; The potential vulnerabilities of the website are easily attacked by viruses, Trojans and malicious programs, resulting in large-scale leakage of personal information [2], etc. These will bring huge economic losses to society and individuals. In recent years, the research of anomaly detection has attracted more and more attention from academia and industry and has been applied to the fields of anomaly detection of big data platforms, machine fault detection, disease detection, identity identification, and fraud detection of credit card or insurance [3]. Under such severe circumstances, it is increasingly urgent to take measures to prevent the deterioration of the network's ecological environment. However, it is impossible to ensure that network crimes never happen. Only by taking effective measures can network anomalies be detected and dealt with in time, which is called network anomaly detection. At present, there are two kinds of network anomaly detection technologies: First, based on the host. This technology mainly detects host logs and operation instructions. Disadvantages: only "after-the-fact" investigation can be realized, but real-time detection is not possible. Internet behaviors are classified into "normal" and "abnormal" behaviors. However, the accuracy of the original anomaly detection method is not high, especially in today's big data era. For the processing of massive network data streams, the execution time is too long to keep up with the needs of the times. Under such circumstances, it is necessary to propose new technologies to improve the efficiency of detection. Therefore, some scholars put forward an anomaly detection method based on behavior analysis. This method studies the internal relationship of the network and tries to consider all aspects of behavior detection. However, at present, some aspects of network abnormal behavior detection technology are still lacking. The focus of detection is on user behavior, and all network behavior patterns are not comprehensively analyzed. Moreover, users are easily influenced by the surrounding environment, and their behavior is unstable, which will interfere with the detection results.

In recent years, with the rapid development of network technology, the network has been an inseparable part of people's daily life. Internet does bring a lot of convenience to users, but

an attack on the internet becomes much more than before. Although many organizations and governmental corporations have established relatively secure protection mechanisms, the means of attack become various and the consequences are much more serious. Under this circumstance, detection and research on internet abnormal behaviors have been gradually developed. After researching several currently mature detective techniques of Internet abnormal behaviors, we find that these techniques are still one-side, detection focuses on users' behaviors, not comprehensively analyzing all Internet behavior modes. Moreover, users are easily affected by their surroundings; behaviors are unstable, which will disturb the result of the detection.

Faced with the above problems, we must seek new methods to effectively identify and deal with network anomalies, and anomaly detection technology based on behavior analysis came into being. It processes network data from the perspective of behavior, which can fully consider the relationship between internal factors, establish network behavior patterns, and thus detect and discover network anomalies. This method makes a deeper study of network behaviors and can provide support for anomaly handling mechanisms. At present, anomaly detection based on behavior analysis has become a research hotspot in the field of network security. However, at present, some aspects of network anomaly detection technology are considered one-sided, and the focus of detection is on user behavior, and all network behavior patterns are not comprehensively analyzed. Moreover, users are easily influenced by the surrounding environment, and their behavior is unstable, which will interfere with the detection results.

Aiming at the abnormal characteristics of big data platforms, this paper mainly detects the system logs and network traffic logs, analyzes the abnormal situation of big data platforms from the network layer and application layer, and ensures the security of big data platforms. For example, the unexpected events on the big data platform and DDOS attacks on the application layer, etc., the traditional academic methods, which use the expert experience to build a rule base, can detect anomalies well when the data volume is small and the attack mode is single before, but there will be some limitations in the era of big data. Given the complexity of modern network attacks, it is difficult to completely detect the abnormal events on the big data platform by the method of building a rule base through expert experience. Moreover, in the traditional non-distributed environment, the analysis and anomaly detection of massive network data will greatly consume computing resources and human resources. Therefore, it is a great challenge to accurately analyze and detect abnormal events in a reasonable time by using data mining [3][4][5], big data [6][7], artificial intelligence [8][9][10], and other technologies.

2. Related Work

2.1. Research status

Intrusion detection includes anomaly detection and misuse detection. Anderson of the University of Pennsylvania first put forward the concept of intrusion detection in 1980. Anderson et al. put forward and implemented an intrusion system by analyzing all kinds of abnormal attacks faced by all kinds of computers, which can well detect the anomalies in the system platform, and the academic community entered the research of intrusion detection. This is also the earliest intrusion detection system model. It can judge the abnormal situation by adopting some relative strategies for data problems. The structure is shown in Figure 1 below.

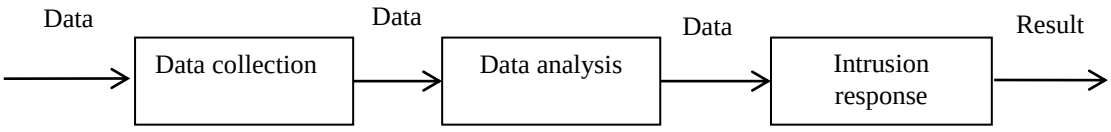

Figure 1: Early intrusion-detection model

Denning et al. proposed an intrusion detection model, which is independent of any known intrusion detection model. Moreover, the model defines the important components of the intrusion detection system in detail, which lays the foundation for future research on intrusion detection systems. Until 1998, faced with a wide range of complex threats and attacks on the network, Tung et al. proposed a general intrusion detection framework that allowed various intrusion systems and various response systems to cooperate. This framework is shown in Figure 2.

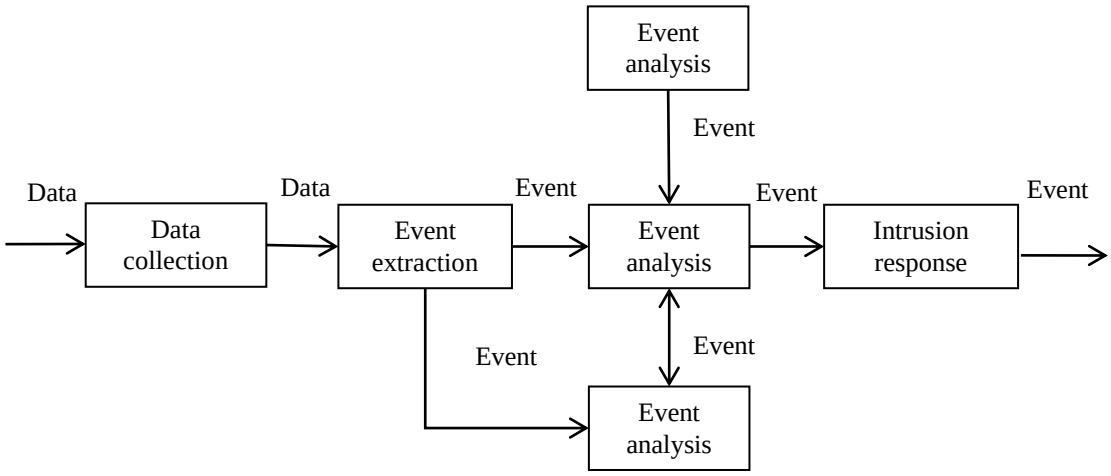

Figure 2: Universal intrusion detection framework

From the perspective of behavior analysis, this paper analyzes and studies the detection of network anomalies. The research mainly focuses on traffic behavior and protocol behavior anomalies, so this paper mainly expounds on the current research status of these two aspects at home and abroad.

(1) Abnormal traffic behavior detection

A single network data can't reflect the user's behavior, and only the data within a certain period can be representative, which can be more accurate and tend to the reality. Therefore, it is necessary to study the network data in a period, that is, traffic. Marina Thottan et al. of bell laboratory proposed a network traffic anomaly detection method based on MIB (Management Information Base) variable mutation analysis and observed the network traffic mutation in MIB variables to judge the occurrence of abnormal traffic behavior. In order to identify

whether there was an anomaly, S.S.Kim et al. retrieved the destination IP address from the packet header, turned it into a discrete wavelet transform, and then performed a statistical analysis. L. American researchers Li et al. suggested using wavelet analysis to compute the energy distribution and use that information to identify DDoS attacks. The so-called energy distribution is a parameter that indirectly reflects whether the flow rate changes. If there is no abnormal phenomenon in the flow rate, the parameter value distribution will be relatively stable. On the contrary, it means that there is an abnormality.

(2) Protocol behavior anomaly detection

It is necessary to detect anomalies from the perspective of the protocol layer. Nowadays, the network data information is encapsulated by the rules of the TCP/IP layer to transmit data, that is to say, all the important data are in the data packet, and each layer of the protocol has its regulations. Only by carefully studying the protocol information of each layer and assisting in detecting abnormal behaviors can we get twice the result with half the effort. At present, from the perspective of the protocol layer, anomaly detection mainly includes: first, based on finite state machine; second, based on the Markov model. Seon Yoon established the TCP protocol state machine model, marked the state of the protocol operation, and then judged whether the TCP protocol operation was abnormal according to whether the occurrence and transition of the state followed the conditions and rules. Kumar Das also judges whether there is any abnormality by whether the state switching is normal. But he created the definition of protocol anonymous detector. Kong Donglin and others in China have established a protocol state machine, which uses regular expressions to match. If there is no match, an exception may be generated.

2.2. Network abnormal behavior detection

2.2.1. The process of network abnormal behavior detection

Generally, the process of network anomaly detection consists of three steps: data collection, model building, and detection. The first step is to screen the data. The data collected initially is redundant and messy, and we need to get the information we need through processing. The second step is to train based on the data processed in the first step, because at first there is no analogy to distinguish between normal behavior and abnormal behavior, so it is necessary to train first. After training a large amount of data, a more accurate model will be obtained, which will help to match the data to be detected in the next step. The third step is to detect anomalies. At present, many behavioral analysis techniques are often used for network anomaly detection. The following section will briefly introduce the analysis methods of network abnormal behavior.

2.2.2. Method of detecting abnormal network behavior

(1) Methods based on probability statistics

This method is the earliest and most widely used detection technology. In the era of big data, the information of network data is huge. Although it is huge, the collection of data will help to analyze the regular information contained in it. The acquisition of behavior is obtained through the statistics of massive data [7]. When we get the user's online law, we match it with the normal law in the database. If it doesn't match the normal behavior law, we think that there is an abnormal situation.

(2) Method based on machine learning

First of all, through the normal data on the network for simulation training, then learn the characteristics of abnormal behavior. The main learning methods are as follows.

a. Inductive learning adopts some data-intensive empirical methods for inductive learning.

b. Analytical learning starts with a few examples and uses domain knowledge for analysis. Its main features are as follows: the reasoning strategy is mainly deduction rather than induction; Use of previous solving experience to guide new problem-solving.

c. Analogy Learning Analogy learning is learning by analogy with specific examples of past experiences.

d. A genetic algorithm is a possible problem. As a vector, each element of the vector is called a gene [11]. The objective function is used to evaluate each individual in the population, and the new population is obtained by the genetic operation of the individual according to the evaluation value.

(3) Method based on data mining

Data mining refers to the process of extracting hidden and useful information from massive data.

(4) Method based on neural network

A neural network is a mathematical model of distributed parallel information processing algorithm that imitates the behavior characteristics of the animal neural network [12][13]. It is an important method to deal with nonlinear systems. At present, many attacks are operated by different attackers on the network, which requires the ability to deal with a large amount of nonlinear data in anomaly analysis.

2.3. Network user behavior

Before analyzing and exploring the network, we have to do some preparatory work. First of all, we have to figure out what kinds of online behaviors are. When we classify, we can start from the purpose of research, the purpose of the behavior, the means of behavior, and so on. The details are as follows.

(1) According to the different transactions of network users using the Internet: the transactions operated on the Internet can be roughly divided into resource search, social interaction, games and entertainment, commodity trading, office work, etc.

(2) According to whether the purpose of Internet users is normal or not: some Internet users use the Internet purely for entertainment or business purposes, but some Internet users use the Internet to attack government or enterprise websites, steal confidential information, invade other people's privacy, and illegally obtain other people's property. According to these two behaviors, Internet behaviors can be divided into normal behavior and abnormal behavior.

(3) According to the number of participating network users, there are four situations: one-to-one, one-to-many, many-to-one, and many-to-many.

One-to-one: "One" of the former refers to network users, and "One" of the latter refers to a specific network. In this case, if the user visits the network frequently, it means that the network user likes the network more, and if the frequency is low, it means that the user doesn't need or like the network so much.

One-to-many: The front "one" refers to network users, and the back "many" refers to some websites. There must be more than one website visited by network users in a certain period, and the number of visits to each website in this period will have a high-low ranking. According to the number of visits to websites by network users, we can judge which websites network users are interested in and which websites are not.

One-to-one: This situation is a hot research topic at present. At present, many shopping websites such as Taobao and Dangdang are warmly welcomed by everyone. In this case, the characteristics of the industry, the characteristics of goods, and personal attributes can be described by collecting data flow, visits, and so on.

Many-to-many: a variety of behaviors of group users.

3. Anomaly Detection Algorithm based on GRU

This chapter mainly analyzes and detects the traffic data collected in the anomaly detection of big data platforms, and constructs relevant detection algorithms. It is mainly from the perspective of the network layer to ensure the security of the big data platform. Entering the era of big data has brought a lot of traversal to people. However, it is estimated that by 2019, the frequent occurrence of global cybercrimes will bring up to $2 trillion in economic losses to global enterprises and people. The big data platform collects a large amount of high-dimensional normal user access traffic data. The past literature demonstrates excellent potential for evaluating and utilizing extensive historical traffic data to successfully ensure the system's dependability and security. It is very important to understand how to rapidly create anomaly detection models and apply them to the anomaly detection analysis of big data platforms using these large-scale traffic data sets with high correlation between distinctive variables. Academics and industry often use data mining and neural networks to discover hidden rules in large-scale historical network traffic data. The most common way to discover intrusions is to analyze user activities [13]. In the literature, Sung et al. proposed a method of network traffic anomaly detection by combining a Support Vector Machine (SVM) with Artificial Neural Network (ANN). In data mining, different data points can be separated by SVM using a hyperplane. Alalshekmubarak et al. put forward a method to classify time series by combining neural networks with SVM in the literature. The algorithm in this paper combines the Echo State Network (ESN), a variant of Recurrent Neural Network (RNN)) with SVM. Nga Nguyen Thi et al. in 2017 [14], proposed a better classification of time series by using the Long Short Term Memory neural network (LSTM) structure and got an ideal result.

The most popular technique for identifying unusual web log data is to examine subscriber behavior patterns. This approach needs a lot of labor and materials. Consequently, it is essential to apply machine learning techniques. A neural network offers significant research and development value and is highly effective in analyzing network logs. A neural network is typically used to examine network logs. Although every node in each layer of the neural network model is unconnected, adjacent layers are connected fully. The method treats each input data separately, making it difficult for it to effectively analyze time-series data like log sequences. The RNN is typically used to examine the prediction series data because the time series properties of the network log data. Suarez-Leon et al [13] propose a method of

classifying network logs by combining RNN with SVM, which can classify network logs successfully but needs a lot of processing power and neurons. GRU is an RNN model with a special structure. The approach can effectively handle the input of time series relations by adding certain connections between hidden layer nodes, and utilizing the GRU unit to regulate the output of data may effectively express the variations of time series. Compared with other RNN models, the active degree of the reset gate can be used to express the length dependence learned by the algorithm. The algorithm's structure makes it clear that GRU is a helpful feature that may effectively extract time series data and resolve the "long dependence problem" in real-world applications. The interior of the GRU deep neural network model is shown in Figure 3.

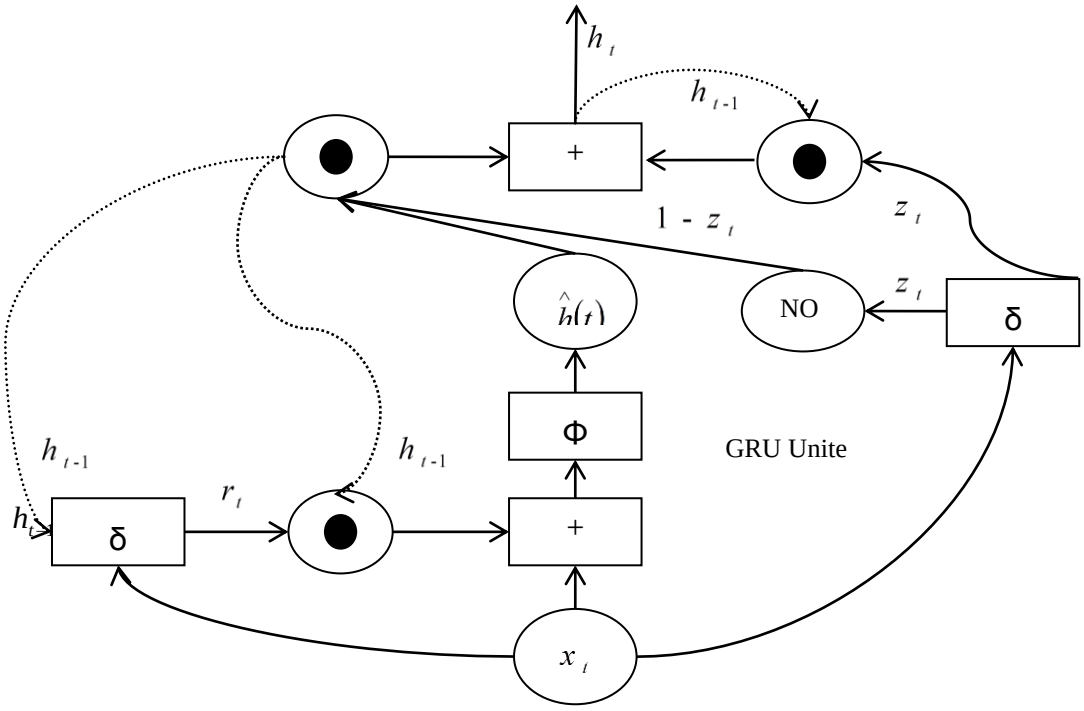

Figure 3: The interior of the GRU deep neural network model

In Figure 2, the dotted line can indicate the hidden node value at the activated t-1 moment, and the filled arrow indicates that it needs to be multiplied by the weight. The h_i shows the activation value of the output. r_t is a reset door, z_t is an updated door, and the formula is as follows.

$$Z_t = \delta\left(W_{zx}x_t + W_{zh}x_{t-1}\right) \tag{1}$$

$$r_t = \delta\left(W_{rx}x_t + W_{rh}x_{t-1}\right) \tag{2}$$

$$\hat{h}_t = \delta\left(W_{hx}x_t + r_t W_{hh}x_{t-1}\right) \tag{3}$$

$$h_t = \left(1 - z_{t-1}\right)h_t + \hat{z}_i\, h_{t-1} \tag{4}$$

Among them, x_t is the current input of the neural network, the output activation value of the last hidden node, δ is the sigmoid function used, the tanh function, and W is the weight matrix. As can be seen from the figure, when the value in the model is close to 0, the information of hidden nodes will be ignored, and only the current moment input information will be used as input. This ensures that the model can lose some useless information to the model. Zt controls whether the information held at the previous moment needs to be brought into the current hidden state. If it is larger, it means that the information provided at the previous moment is more.

The abnormal examination of the network logs on the big data platform, which include extremely unbalanced internal data, is the topic of this paper. The majority of the offline network log files collected is just regular data. To evaluate this type of network log, picking a strong classification model is more beneficial. The output layer of a traditional neural network uses Softmax to classify network logs. The derivative value of the Softmax function is between 0 and 1, as shown in Figure 3. The number of network layers is inversely proportional to the error of output layers, that is, the gradient becomes smaller, and at the same time, the phenomenon of "gradient disappearance" is brought about, which is one of the reasons for the wrong classification of the traditional model.

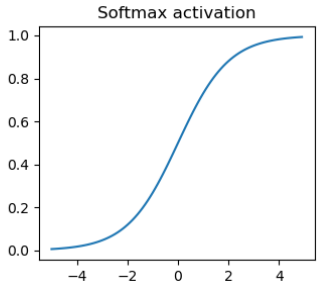

Figure 4: Sigmoid function image

To facilitate the description of vector data, Tax and Duin proposed a new single classification approach called SVDD (Support Vector Domain Description) in 1999. This method is based on the idea of Support Vector Machines (SVM). One of the most popular support vector learning techniques is this one. A ball defined on the feature space is used by the algorithm to attempt to separate a collection of normal data from all other potential abnormal things. If it is necessary to judge that there is only one class, SVDD will include the log data of this class by finding the most suitable hypersphere. Input the data to be identified into the SVDD model. If the data falls into the hypersphere, it is determined that the data belongs to this class. Otherwise, the data does not belong to this class. Therefore, SVDD can solve this problem better. In this paper, an anomaly detection method based on improved GRU is proposed, and the GRU output is improved by the SVDD algorithm. First, to extract the positive sample set X from the original network traffic data set, which contains N positive sample data, (i = 0,1, 2,…, N) quantity, and other parameters, we assume that the normal network access traffic can surround the spherical center of the hypersphere, with the spherical center α and the radius R, and then we can determine an optimization equation as follows:

$$\min\left|R^2 + C\sum_{i}^{N}\varepsilon_i\right| \tag{5}$$

$$s.t. \quad R^2 = \varepsilon_i - (X_i - \alpha)(X_i - \alpha) \geq 0 \tag{6}$$

Formula (1-6) is a quadratic programming problem. By introducing Lagrange multiplier into the formula to improve this method, Formula (7-8) can be modified as follows:

$$\max L = \sum \alpha_i (X_i \bullet X_j) - \sum \alpha_i \alpha_j (X_i \bullet X_j) \tag{7}$$

$$s.t. \quad \sum \alpha_i = 0, if(\alpha_i \geq 0) \tag{8}$$

Where $\|z - \alpha\|^2 \leq R^2$ is the support vector. For the test containing abnormal sample Z, when the sample satisfies formula (1-8), that is, it is normal traffic access, otherwise it is abnormal.

4. Experimental Design and Result Analysis

The improved GRU anomaly detection algorithm is developed in Python within the Windows operating system to demonstrate its feasibility. The development environment consists of a laptop with a 4-core CPU and 16G of memory, while PyCharm is the development tool. In the experiment, to realize the algorithm proposed in this section efficiently, Tensorflow1.6, an open-source machine learning library provided by Google, will be used to build an available GRU neural network model.

4.1. Data set

In this study, the KDD CUP 99 data set—among others—is chosen and created for the network traffic anomaly detection method's application. One of the few publicly accessible data sets, it is extensively used in the network-based anomaly detection system. The single record's 42 available statistical features include class marking, which classifies a range of aberrant and normal types and stamps them all as normal. The four most frequent attack types in this study are DOS, R2L, U2R, and PROBE. 10% of the KDD CUP 99 data set is chosen for this experiment in order to train the model. Moreover, in this section, the experimental data set KDD CUP99 will be randomly selected and divided into five groups for effective comparison in many aspects.

In the anomaly detection of network traffic data, this section uses two performance indicators: Detection Rate (DR) and False Alarm Rate (FAR) to evaluate the anomaly detection effect of the model. The ratio of abnormal users (TP+TN) to all abnormal users (P+N) represents the detection rate, whereas the ratio of abnormal users (FP+FN) to all normal samples (P+N) represents the false alarm rate. The experimental design in this section is applied to KDD CUP 99 data set. Since the second, third, and fourth dimensions of the data set are not numerical, they cannot be entered directly into the model of the anomaly detection algorithm. The three-dimensional data will undergo numerical preprocessing before to the experiment. It is possible to replace the original content by counting how often each of these three features will appear initially, after which they can be sorted by letter or serial number. The first group of experiments in this section is to select the appropriate number of principal components to improve the efficiency of anomaly detection. When choosing different numbers of new principal components, the number of new principal components is the optimal choice of principal components when the anomaly detection efficiency is at its

highest. This section's second set of studies compares the algorithm's detection rate and false alarm rate to those of various recently developed techniques. The first comparison is the method of this section with the classical GRU-MLP [14], LSTM [15], PCA-SVM [16] and LSTM-RNN [17]. GRU-MLP, LSTM-RNN, and LSTM are classic anomaly detection methods, which can quickly detect anomalies in the network, but their detection efficiency is low when applied to large-scale data sets. PCA-SVM is an efficient method to improve SVM, but this method simply reduces the dimensions of large-scale data sets, and these algorithms are all classification methods that need all class labels, and can only detect known abnormal situations.

The algorithm suggested in this paper is an improvement of the GRU algorithm, which uses both an effective SVDD single classification algorithm to replace the output of GRU and PCA to reduce the dimension of data. As a result, the algorithm is able to identify unknown anomalies with high detection sensitivity.

This experiment is also applied to a large number of parameters. The experiment will be based on the research of Suarez-Leon et al. [13], including that the algorithm parameters with multiple neural networks in the algorithm are adjusted manually, and are not automatically optimized through super-parameters, as shown in Table 1.

Table 1: Hyperparameters in neural networks

Super parameter	GRU-SVDD
Batch Size	256
Cell Size	256
Dropout Rate	0.85
Epochs	five
Learning Rate	1*10-5

4.2. Experimental analysis

The training set for this approach randomly chooses 30,000 pieces of traffic data from the attribute label Normal to train a single classification model on. The test data set is divided into five groups to test the detection efficiency of the algorithm, as shown in Table 2. The first group is a test data set containing four abnormal situations, the second group contains only DOS attacks, the third group contains R2L attacks, the fourth group contains U2R attacks, and the fifth group contains PROBE attacks.

Table 2: Data grouping

data clustering	Normal	DOS	R2L	U2R	PROBE	total
the first group	2000	2000	1000	500	1500	6000
the second group	2000	3000	0	0	0	5000
Third groups	2000	0	1000	0	0	3000
Fourth groups	2000	0	0	500	0	2500
Fifth groups	2000	0	0	0	1500	3500

The number of principal components chosen by characteristics following data processing using principal component analysis is the subject of the investigation. In order to demonstrate the impact of data processing using principal component analysis on detection rate, the

experiment chooses the first batch of data sets, which includes a variety of anomalous samples. In Figure 3, the ordinate shows the algorithm's detection rate, while the abscissa shows how many main components were chosen via principal component analysis.

To compare the detection of anomaly detection model algorithms to detect these four kinds of anomalies, the classical BGRU-MLP, LSTM [15], PCA-SVM [16], and LSTM-RNN [17] are selected for comparison with the improved GRU algorithm in this section. The detection rate and FAR of the four algorithms are compared in the above five groups of experiments.

Table 3: Detection of four types of anomalies by four algorithms

type	index	mix	DOS	R2L	U2R	PROBE
GRU-SVDD	Detection rate	98.7%	99.6%	98.5%	56.3%	96.7%
	False alarm rate	2.4%	0.01%	5.2%	0.09%	0.05%
GRU-MLP	Detection rate	98.2%	99.80%	50.40%	53.84%	95.3%
	False alarm rate	0.82%	0.08%	0.03%	0.05%	0.30%
LSTM-RNN	Detection rate	97.6%	99.2%	3.06%	4.9%	53.2%
	False alarm rate	6.5%	0.9%	0.03%	0.012%	0.14%
LSTM	Detection rate	77.7%	99.5%	92.9%	60.0%	84.7%
	False alarm rate	7.2%	0.7%	0.04%	0.02%	0.2%
PCA-SVM	Detection rate	96.9%	100%	97.50%	17.54%	95.60%
	False alarm rate	5.88%	0.18%	14.73%	0.40%	0.20%

It is clear from Table 3's experimental results that the improved GRU anomaly detection model put forth in this section may be successfully used with traffic data sets. It is clear that the approach in this part has a higher detection rate for four different types of anomalies when compared to the traditional GRU-MLP, LSTM-MLP, LSTM, and PCA-SVM techniques. Additionally, the other two systems' detection efficiencies for U2R and other anomalous circumstances are quite poor, making the approach suggested in this study preferable to them. However, in terms of FAR, the algorithm suggested in this research is unstable. Good precision exists for two reasons. In order to reduce the redundant characteristics in the original records, the algorithm in this research first employs PCA to reduce the attribute dimension of network traffic data sets. Second, the GRU neural network used in this paper only uses regular network traffic to build an effective single classification model called SVDD. As a result, it is capable of checking unusual circumstances like U2R and can be used to identify unidentified attacks. To further compare with more classical algorithms, this paper uses the first set of data sets under mixed conditions to compare this algorithm with the detection rates of BP neural network, Semi-Supervised GHSOM, PCA-BP, GRU-Soft Max, BGRU-MLP [19], as shown in Table 4.

Table 4: Six algorithms for detecting mixed anomaly types

Algorithm	Detection rate
GRU-SVDD	98.7%
BP	69.5%
PCA-BP	93.3%
Semi-Supervised GHSOM	91.2%
GRU-SoftMax	70.75%
BGRU-MLP	98.2%

Table 4 shows that the modified GRU-based anomaly detection algorithm successfully detects anomalies in network traffic data sets with high correlations between high dimensions and characteristic features. Four popular and straightforward deep learning methods are compared at the same time. They all create categorization models for sets of network traffic data using neural network technology. These algorithms, however, do not account for unidentified attacks and instead classify the data sets according to a variety of normal and abnormal attributes. The usual network traffic is modeled in this article; anything outside of the normal range is considered abnormal. Additionally, this is reliant on the features of the big data platform. A single classification model can be updated based on identified traffic increases because the majority of the access traffic that big data platforms collect is normal. The algorithm suggested in this research is also better than the other five algorithms in terms of detection rate.

5. Conclusion

The classic approaches for network traffic anomaly detection are presented first in this study. They have low detection efficiency for novel and unidentified assaults and cannot be effectively applied to large-scale and high-dimensional historical access traffic data sets on big data platforms. This paper proposes an anomaly detection technique based on enhanced GRU. First, principal component analysis is used to preprocess the high-dimensional raw data set in order to remove duplicated features and boost detection effectiveness. The properties of the preprocessed network traffic time series data are then extracted using the GRU neural network structure. The next step is to represent the typical network traffic during classification and output using SVDD, an effective single classification technique. This approach is more useful and effective. The algorithm based on modified GRU has higher accuracy of anomaly detection than the conventional technique, as demonstrated by a large number of experiments on the experimental data set and the analysis of the experimental results. The algorithm based on the enhanced GRU put forth in this research is not without flaws. The experiment solely uses principal component analysis to lessen attribute redundancy, which increases the algorithm's detection performance. The algorithm first adds a principle component analysis to preprocess the original data set. However, no principal component reduces the number of attributes, which may be more efficient in time. Secondly, the experiment in this paper uses KDD99, an open-source data set. The algorithm proposed in this paper realizes the anomaly detection function of user behavior, including the anomaly detection model for two types of logs. The system can match different background processing methods according to the log types uploaded or monitored by users.

Aiming at the system logs to be processed by users, the anomaly detection method based on sequential pattern mining should be used in the system. Firstly, the method needs to build a rule base according to offline normal user access log information, including data cleaning of Weblogs and deleting irrelevant and redundant information; User identification, identifying all access information of each access user; Session identification, which separates a series of access link sets with time relationship from a user's visit to a certain server to his departure for subsequent pattern mining; Pattern mining, using the improved distributed maximum frequent sequence pattern extraction algorithm to extract user access patterns, that is, rules. Then, the session sequence is extracted from the data to be detected by the user, and the sequence is compared with the rules to detect and locate the abnormal situation. Aiming at the traffic logs to be processed by users, the system should adopt the anomaly detection method based on improved GRU. Firstly, the method needs to be constructed according to the offline

traffic log data selected by the user. It is proved that this data set is related to the data set involved in this paper, but it is not the same. It may be unsuitable in practical application, and some experimental parameters need to be adjusted, which is not well explained.

Reference

[1] Praseed, A. & Thilagam, P. S. (2018). DDoS attacks at the application layer: challenges and research perspectives for safeguarding web applications. *IEEE Communications Surveys & Tutorials*.

[2] Mehta, S., Kothuri, P., & Garcia, D. L. (2018). Anomaly detection for network connection logs ar Xiv preprint ar Xiv:1812.01941.

[3] Javier, L. Z., Torre Albo, J., & Cristobal R. (2021). Early prediction of student learning performance through data mining: A systematic review. Psicothema, (2021), 33(3), 456-465.

[4] Nguyen, T. V., Zhou, L., & Chong, A. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*, 281(3), 543-558.

[5] Peji, A. & Molcer, P. S. (2021). Predictive machine learning approach for complex problem-solving process data mining. *Acta Polytechnica Hungarica*, 18(1), 45-63.

[6] Davoudian, A. & Liu, M. (2020). Big data systems. *ACM Computing Surveys (CSUR)*.

[7] Gronemeyer, H. & Souren, N. Y. (2021). Big data: The good, the bad, and the ugly. *International Journal of Cancer*.

[8] Nina, S. & Brian, W. (2020). Artificial intelligence and the future of global health. Lancet (London, England), 395(10236), 1579-1586.

[9] Mamdani, M. & Slutsky, A. S. (2020). Artificial intelligence in intensive care medicine. *Intensive Care Medicine*, (5).

[10] Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management*, 37(3), 212-227.

[11] Tabak, B. M., Silva, T. C., & Zhao, L. (2020). Applications of machine learning methods in complex economic and financial networks, *Complexity*.

[12] Wen, G., Wen, P., & Tang, Z. (2021). Research on data mining method of TCM prescription based on machine learning. *Journal of Physics: Conference Series*, 1952(2), 022033.

[13] Suárez- León, A. A., Varon, C., & Willems, R. (2018). T-wave end detection using neural networks and Support Vector Machines. *Computers in Biology & Medicine*, 96, 116.

[14] Thi, N. N., Cao, V. L., & Le-Khac, N. A. (2017). One-class collective anomaly detection based on LSTM-RNNs.

[15] Zhao, Y., Jie, L., & Shuang, X. (2017). Investigating gated recurrent neural networks for acoustic modeling. *International Symposium on Chinese Spoken Language Processing*.

[16] Agarap, A. F. M. (2018). Proceedings of the 2018 10th International Conference on Machine Learning and Computing, - ICMLC 2018 - A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data.

[17] Anki, P., Bustamam, A., & Al-Ash H. S. Intelligent chatbot adapted from question and answer system using RNN-LSTM model. *Journal of Physics Conference Series*, 1844(1), 012001.

[18] Kim, J., Kim, J., & Thu, H. L. T. (2016). Long short-term memory recurrent neural network classifier for intrusion detection. *2016 International Conference on Platform Technology and Service*, IEEE.

[19] Yin, C., Zhu, Y., & Fei, J. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*, 5(99), 21954-21961.

This page is empty by intention.