Research on Intrusion Detection Method Based on Hierarchical Self-convergence PCA-OCSVM Algorithm

Yanpeng Cui, Zichuan Jin, and Jianwei Hu
(Corresponding author: Zichuan Jin)

College of Network and Information Security, Xidian University
North Campus of Xidian University, Xi'an 710071, China
(Email: 546720018@qq.com)

(Received July 2, 2019; Revised and Accepted Dec. 28, 2019; First Online Apr. 6, 2020)

Abstract

At present, traditional intrusion detection methods have some shortcomings, such as long detection time, low detection accuracy and poor classification effect. This paper will combine PCA and OCSVM algorithm to build a multi-level intrusion detection model, using attack feature analysis method to preprocess data, while data cleaning and data feature selection of training set. It highlights the characteristics of abnormal data and normal data, and weakens the influence of irrelevant features on training model. PCA algorithm is used to process data to improve detection rate and reduce noise. Different models are trained by different data features to detect four attack types, namely Probe, DDOS, R2L and U2R. The optimal dimension of PCA is automatically obtained by calculating the contribution rate M of feature, which improves the traditional method that requires frequent input of K value. The model is trained by using OCSVM algorithm based on RBF core, and the disadvantage of poor classification effect of OCSVM algorithm is eliminated through improved multi-layer detection mechanism. Finally, the KDDCUP99 data set is used for experimental verification. The results show that the proposed method has more advantages than the traditional detection method.

Keywords: Intrusion Detection; KDDCUP99; Self-Convergent PCA-OCSVM

1 Introduction

With the rapid development of computer technology, people pay more attention to information security. According to CNCERT 2018 overview of China's Internet network security situation [11], it is found that there are more serious apt attacks, data leakage, distributed denial of service attacks in 2018. In 2018, CNCERT handled 106000 network security incidents. The main types of security incidents are system vulnerability exploitation and DDOS attacks. Through the combination of basic telecom enterprises and cloud service providers, the DDOS attacks launched in 2018 fell 46% year-on-year, and the accused end fell 37% year-on-year. The defense measures of security experts have played a certain role, but there are still about 20000 government websites implanted in the back door. Therefore, while strengthening the network defense means, the research of intrusion detection also needs to continue in-depth. At present, the traditional intrusion detection method mainly relies on the regular matching method to analyze the structured data stored in the database. For example, the success of Snort [23] and other intrusion detection systems is based on strong prior knowledge and customized attack rule set, but it is difficult to effectively detect unknown attacks. In addition, when the Snort Intrusion detection system matches too many or too complex rule sets, it will have a great impact on the performance of the server itself, reduce the detection rate, and even lead to the collapse of the intrusion detection system. With the rise of the field of artificial intelligence, researchers found that most of the machine learning algorithms can be applied to the field of intrusion detection with appropriate changes based on their mathematical principles [6,12,16,26].Different machine learning algorithms can achieve better results by combining with other algorithms. Intrusion detection by machine learning can reduce the workload of manual data analysis, and find more differences between abnormal data and normal data in the way of digital characteristics. Combined with the big data analysis method, according to the network traffic and the information brought by the log, we can explore the deeper correlation within the security events. To realize intrusion detection methods with higher detection rate, higher accuracy and more types of detection attacks [5].

The one class SVM studied in this paper is a classification of SVM algorithm. In [17] schölkopf and others proposed a class of SVM (one class SVM) algorithm. Its

main principle is to train data set by support vector machine, separate data points from the feature space of the origin, and maximize the distance from the hypersphere to the origin. According to different probability density, a hypersphere is divided, and the data in the area of small probability density is divided into abnormal data. One class SVM usually needs kernel function to solve nonlinear problems. Kernel function can make vector calculate inner product directly in the original low dimensional space, avoiding the complex calculation directly in the high dimensional space. Common kernel functions include linear kernel function, polynomial kernel function, Gaussian kernel function, etc. Among them, Gauss kernel is very flexible and one of the most widely used kernel functions. When using Gaussian kernel function, the choice of its parameters will have a great influence on the formation of hypersphere. In this paper, the principal component analysis (PCA) is improved. By calculating the characteristic contribution rate M, the optimal dimension of PCA is automatically obtained, and the traditional method which needs frequent input of K value is improved. Using the improved PCA algorithm to reduce the dimension and noise of the data set, make the hypersphere generated by one class SVM smooth enough, and reduce the impact of noise points on the hypersphere [19].Finally, the KDDCUP99 data set was used. The data set was produced in 1999 by DARPA, an intrusion detection evaluation project in MIT Lincoln Laboratory. Although the data set was collected in the attack log 20 years ago, it still has important reference significance for the characteristics of current network attacks. At present, although there are many changes in the means of network attack, the traffic caused by the attack is still similar to the information characteristics recorded in the log [21].
The main contributions of this paper are as follows:

1) By using the attack feature analysis method to pre filter the data, at the same time, the training set is cleaned, selected, digitized and normalized. Highlight the characteristics of abnormal data and normal data, and weaken the influence of irrelevant features on training model;

2) The PCA algorithm is improved, and the best dimension of PCA is obtained automatically by calculating its characteristic contribution rate M, which improves the traditional method that needs frequent input of K value;

3) By combining the characteristics of PCA algorithm and OCSVM algorithm, a layered PCA-OCSVM algorithm detection framework is proposed to optimize the detection model;

2 Related Research

At present, the research on SVM algorithm is still hot. Yingchao Xiao *et al.* [24] proposed to use MIES method

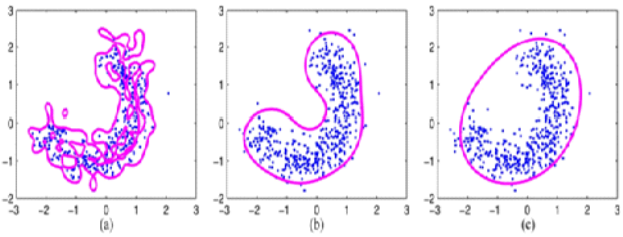

Figure 1: Different gaussian nuclear parameter results

to automatically obtain the optimal Gaussian kernel parameters, so as to obtain the optimal hypersphere. Using different Gaussian kernel parameters will have different effects on clustering results. As shown in Figure 1, this paper uses MIES method to study the geometric position of the edge and internal sample mapping in the feature space relative to the OCSVM hyperplane, and uses the distance difference between the internal sample and the edge sample to the closed surface to evaluate the applicability of the Gaussian kernel parameter D. Through the appropriate evaluation method, the optimal Gaussian kernel parameters are obtained. It can be seen that the selection of parameters of Gaussian kernel function directly affects the final classification effect. Cui Mei Bao [4] proposes to use one class SVM to implement intrusion detection based on SNMP MIB data set. Since the output of one class SVM defined in different feature spaces represents the absolute distance between the corresponding data and the decision boundary, it is not feasible to determine the related classes by comparing the absolute distances in different feature spaces. This method emphasizes the detection of DDOS. According to the protocol, DDOS is divided into tcp-syn flooding, ICMP flooding and UDP flooding. After adjusting the corresponding parameters, the classification results of DDOS attacks are better, all of them are over 98%, and the false alarm rate is still high at 9%.

Ming Zhang *et al.* [8] proposed one class SVM detection method based on Gaussian kernel function, and carried out security analysis on KDDCUP99 data set. At the same time, a new network intrusion detection model based on a class of support vector machines is proposed. The accuracy of using this model to detect normal data is as high as 100%. However, the disadvantage is that the detection rate of R2L and U2R attacks based on a class of support vector machine model is relatively low, only 26.85% and 69.23%, part of the reason that affects the accuracy of the results is the lack of data, and the establishment of R2L and U2L models is not comprehensive.

From the existing research, it can be found that OCSVM based intrusion detection has the problems of low detection rate for low-frequency attacks and single type of detection attacks [25].Therefore, this paper proposes a pca-ocsvm multi-layer detection method, which optimizes the detection effect of OCSVM multi-layer model by preprocessing and feature extraction of KDD-

CUP99 data set and smoothing with PCA algorithm.

3 Algorithm Research

3.1 Self-Convergence PCA

Principal Component Analysis (PCA) is mainly supported by covariance and covariance matrix. In signal processing, it is considered that the signal has larger variance and the noise has smaller variance. By filtering out the signal with smaller variance, the overall signal quality can be improved [10].PCA algorithm is mostly used in image processing and data dimensionality reduction. Through linear mapping, the high-dimensional data vector is projected onto the low-dimensional space, and the main components of the data are retained. That is to say, the data features with large variance are retained, and the unimportant part of data description is weakened. This can not only retain the main characteristics of data, but also reduce the amount of calculation and improve the efficiency of operation. The improved PCA algorithm flow is in Algorithm 1.

Algorithm 1 Working of the self convergence PCA

1: Begin
2: Algorithmic input: Input data set $\mathbf{X}_{mxn}$.
3: Calculate the mean $\mathbf{X}_{\text{mean}}$ of the data $\mathbf{X}_{mxn}$. set $\mathbf{X}_{\text{new}} = \mathbf{X}_{mxn} - \mathbf{X}_{\text{mean}}$
4: The calculated covariance $\mathbf{X}_{\text{new}}$ matrix is denoted as $\mathbf{X}_{\text{cov}}$,and computed eigenvalues and eigenvectors of $\mathbf{X}_{\text{cov}}$.
5: Arrange the eigenvalues from big to small, select the first k values and take the corresponding k eigenvectors as column vectors to form a matrix $\mathbf{X}_{nxk}$
6: Computing $\mathbf{X}_{\text{new}}$ $\mathbf{X}_{nxk}$,the dimension-reduced data set $\mathbf{X}_{\text{new}}$ can be obtained by projecting the matrix composed of the data set to the matrix composed of the selected feature vectors $\mathbf{X}_{\text{new}}$ $\mathbf{X}_{nxk}$.
7: Set the threshold value m according to the contribution value of each dimension of the reduced dimension data set,and round off the dimension that does not reach the threshold value, so that the number of remaining dimensions is p;
8: **while** the contribution rate of a certain dimension is less than m **do**
9: Let $k = p$ and return to step 3 until all dimension contribution values greater than or equal to m.
10: **end while**
11: End

The main components of data set screened by PCA algorithm have the following properties:

1) The principal components are orthogonal, and the difference is more significant;

2) The variance of principal components decreases in turn;

3) The data characteristics after processing lose their original explanatory nature;

4) The total amount of information remains unchanged.

Using PCA algorithm to preprocess data characteristics can highlight the internal differences of data characteristics, reduce the data processing dimension and maximize the characteristics of normal data and abnormal data, which is conducive to further exception analysis of subsequent algorithms [27]. When dealing with data sets with higher dimensions and there is a certain correlation between dimensions, PCA can be used to recombine attributes into uncorrelated principal components to represent the original information. Using PCA algorithm can also effectively reduce the dimension of sample set and improve the efficiency of operation.

3.2 OCSVM

Schölkopf *et al.* [17] extended the original SVM algorithm and proposed OCSVM algorithm. Its core idea is to transform a classification problem into a binary classification problem through hypersphere. Based on known input data sets $D = \{\mathrm{x}_i\}$,$\mathrm{x} \in \mathrm{R}^N$,$1 \leq i \leq n$.At the same time, it is assumed that there is a mapping χ from original space R^N to multidimensional space φ,and $\varphi(x_i) \in \chi$.At this point, the problem is transformed into finding a binary classifier, which divides the high-density region containing most of the normal sample points into some anomalous discrete points, which are recorded as '+1' and '-1'.

In this paper, we mainly use the Gauss kernel as the kernel function of OCSVM. For the Gauss kernel function, there are:

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{s}),$$
$$\langle \varphi(x_i)\ ,\ \varphi(x_j) \rangle = K(x_i, x_j) = 1.$$

It can be found that the training samples are mapped to the feature space and distributed on the circle with coordinate origin as the center and radius $R = 1$. Gaussian kernels can effectively avoid the impact of data standardization and bring a very smooth estimation to optimize the classification effect. Gaussian kernels can adjust the fitting degree by adjusting the scale parameter s.

4 Hierarchical PCA-OCSVM Model

This paper presents an anomaly detection method based on PCA-OCSVM, which integrates the characteristics of PCA and OCSVM. By extracting different data features of KDDCUP99, the original data are digitized and normalized, and then input the data set into the layered detection model [18]. Comparing the performance of OCSVM linear kernel with Gauss kernel in the algorithm, it is found that Gauss kernel has better detection effect in

Table 1: Samples of raw training and testing KDD Cup 1999 dataset

num	Example
1	0,tcp,http,SF,291,1096,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,29, 255,1.00,0.00,0.03,0.05,0.03,0.01,0.00,0.00,normal.
2	0,tcp,http,SF,219,1098,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,7, 255,1.00,0.00,0.14,0.05,0.00,0.01,0.00,0.00,normal.
3	26,tcp,ftp,SF,116,451,0,0,0,2,0,1,0,0,0,0,0,1,0,1,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,1, 1,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,0.00,ftp_write.
4	0,icmp,eco_i,SF,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,100,1,0.99,0.00,0.00,0.00,0.01,0.08,0.00,13, 245,0.23,0.15,0.23,0.25,0.00,0.00,0.08,0.00,ipsweep.
5	0,icmp,eco_i,SF,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0.00,0.00,0.50,0.00,0.50,1.00,0.00,14, 245,0.29,0.14,0.29,0.25,0.00,0.00,0.07,0.00,ipsweep.

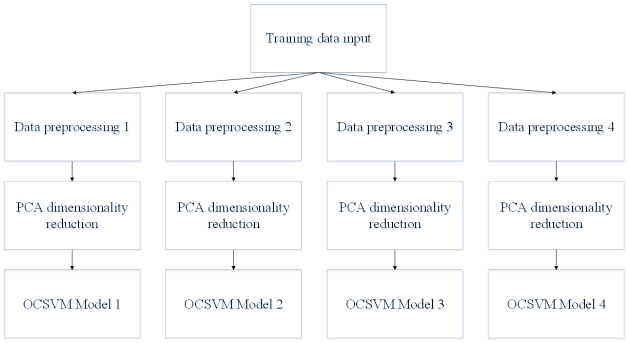

Figure 2: PCA-OCSVM training model

detection because of its superiority in dealing with nonlinear data.

4.1 PCA-OCSVM Training Method

This section mainly puts forward the training method for OCA-OCSVM model, and the method for obtaining the training model is shown in Figure 2. For the data preprocessing method in this figure, the method of digital substitution and normalization is mainly used to preprocess the KDDCUP99 data set. The original data is shown in Table 1. Some replacement methods are given in Table 2. The fields 2,3,4 columns of the dataset are digitized in a way similar to Table 2. The digitized dataset will be identified by PCA algorithm to form a matrix. The results after replacing the original data are shown in Table 3. The function of PCA algorithm is to expand the variance within the data. If the variance of some data features in the data is very large, it cannot highlight the difference between normal data and abnormal data. Therefore, it is necessary to filter and normalize the data features, otherwise the normal data and abnormal data will be confused, and the detection rate will be reduced. In the data preprocessing stage, firstly extract the data features of KDDCUP99 data set. According to the difference between U2R, L2U, Probe, DDOS attack data and normal data in different dimensions, filter the data features with large travel differences to form a new training data set, and then normalize the new data set to reduce the impact of one of the features on the data set. Expand the comprehensive impact of different dimensions of data, improve the detection accuracy. Data normalization is defined as follows:

$$X_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

Table 2: Conversion table

Raw data	tcp	udp	icmp
Replacement data	1	2	3

Then PCA is used to reduce the dimension of the normal data. Finally, we use OCSVM based on Gauss kernel to train the normal model, and get the normal model in four different dimensions.

4.2 PCA-OCSVM Detection Method

In this section, a detection model of multi-layer PCA-OSVM is proposed. For the four different normal models which are trained to be used for anomaly detection of test data sets, the anomaly detection flow is shown in Figure 3.

1) The DDOS detection model is used to determine whether a DDOS attack is established or not, and the data not considered as a DDOS attack is transferred to the next model.

2) The probe detection model is used to determine whether the probe attack is valid or not, and the data not considered as the probe attack is transferred to the next model.

3) The R2L detection model is used to determine whether the R2L attack is valid or not, and the data not considered as the R2L attack is transferred to the next model.

4) U2R detection model is used to judge whether U2R attack is established or not, and data not considered as U2R attack is regarded as normal data.

Table 3: Data samples after KDDCUP99 digitization

num	Example
1	0,1,22,10,291,1096,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,29, 255,1.00,0.00,0.03,0.05,0.03,0.01,0.00,0.00,normal.
2	0,1,22,10,219,1098,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,7, 255,1.00,0.00,0.14,0.05,0.00,0.01,0.00,0.00,normal.
3	26,1,21,10,116,451,0,0,0,2,0,1,0,0,0,0,1,0,1,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,1, 1,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,ftp_write.
4	0,3,24,10,10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,100,1,0.99,0.00,0.00,0.00,0.01,0.08,0.00,13, 245,0.23,0.15,0.23,0.25,0.00,0.00,0.08,0.00,ipsweep.
5	0,3,24,10,10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0.00,0.00,0.50,0.00,0.50,1.00,0.00,14, 245,0.29,0.14,0.29,0.25,0.00,0.00,0.07,0.00,ipsweep.

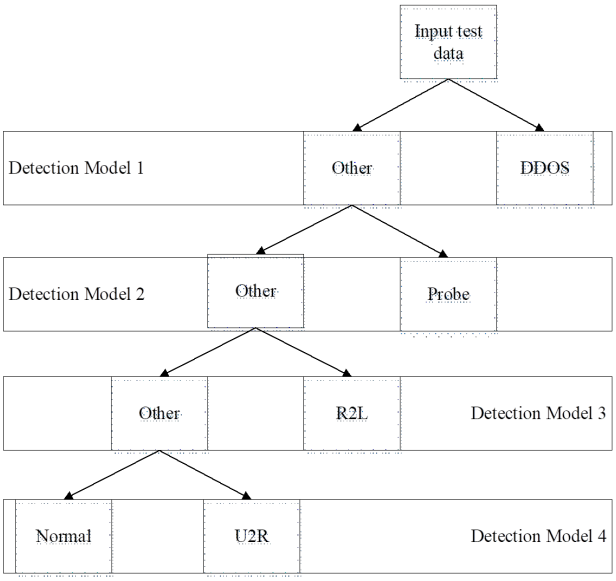

Figure 3: PCA-OCSVM detection model

OCSVM algorithm can only detect one model, and for using different kernels, the detection results will be very different. By analyzing the characteristics of KDDCUP99 data set and different kernel functions of OCSVM, this paper finds that OCSVM based on linear kernel functions performs poorly in detection results, and it is difficult to distinguish abnormal data and normal data, while using Gaussian kernel OCSVM based on different kernel parameters gets better detection results [13]. Through the layered detection mechanism, the disadvantages of OCSVM which can only detect one model are improved, and the advantages of OCSVM in solving unbalanced sample classification are continued. On the basis of making OCSVM as simple and fast as possible, it can detect many kinds of exceptions and expand the available scenarios of the algorithm.

4.3 Preprocessing of Training and Test Data

In this section, KDDCUP99 data set will be analyzed and feature selection [15], and irrelevant features in different models will be cleaned to improve the identification of detection model. The original KDDCUP99 data set contains 41 dimensional data characteristics, including the basic characteristics of TCP connection (9 kinds in total), the content characteristics of TCP connection (13 kinds in total), the time-based network traffic statistical characteristics (9 kinds in total), and the host based network traffic statistical characteristics (10 kinds in total). However, for different attack features corresponding to different attack detection models, under our proposed layered detection model, redundant features will interfere with the correct training of the model [7]. Therefore, it is necessary to pre filter the data set. According to the differences of data records caused by different kinds of attacks, the data dimensions related to the attack principle are saved and trained, and the data features irrelevant to the corresponding attack types in the 41 dimensional features are screened. For example, the data phenomenon caused by scan attack comes from the basic characteristics of TCP connection and the statistical characteristics of host network traffic. The content characteristics of the scan attack for TCP connections are roughly the same as the normal traffic, so the data samples filtered according to the scan attack characteristics are shown in Table 4. In this section, by analyzing the characteristics of Probe, DDOS, U2R and R2L attacks, different attack features are extracted to test the intrusion detection model, so as to highlight the data set differences caused by different attack modes and improve the detection accuracy [20, 22]. The data feature dimensions retained after filtering are shown in Table 5.

5 Experiments and Results

In this section, we will test the proposed layered PCA-OCSVM detection model on the VM virtual machine of Ubuntu 16.04, using 3G memory, virtual machine with 4-

Table 4: Porbe replaced data samples

num	Example
1	0,10,222,773,0,11,11,0.00,0.00,0.00,0.00,normal.
2	0,10,212,786,0,8,8,0.00,0.00,0.00,0.00,normal.
3	0,10,260,1837,0,11,11,0.00,0.00,0.00,0.00,normal.
4	1,5,0,0,0,128,2,0.00,0.00,0.53,1.00,portsweep.
5	0,6,0,0,0,17,1,0.05,1.00,0.02,0.00,ipsweep.

Table 5: Data feature dimension after filtering

Attack types	Data Feature Dimensions Retained after Screening
Probe	1,4,5,6,11,23,24,38,39,40,41
DDOS	5,6,13,23,24,25,26,29,30,32,33,34,35,37
U2R	1,3,10,11,13,14,15,16,17,18,19,23,24,25,31,32,33,34,35,36
R2L	4,10,11,14,15,16,17,18,19,23,24,27,28,32,33,36,38,39,40,41

core CPU performance and python 2.7.6 compilation environment. KDDCUP99 data set is adopted for training and test data. After analyzing contribution rate of each dimension by improved self-convergence PCA algorithm, the threshold value of contribution degree is m = 0.001. Dimension parameters under different attack models obtained by self-convergence are shown in Table 6. Table 7, Table 8, Table 9 and Table 10 show the detection results of intrusion detection test based on the parameter model obtained by self-convergence algorithm. The results show that this method can optimize the engineering efficiency, quickly obtain excellent dimensional parameters, and has a high detection accuracy. Let $j\epsilon\{Probe, DDOS, U2R, R2L\}$, i be the corresponding subclass attack under each big class attack. The accuracy is AC_{ji}, the number of tests is TQ_{ji}, the number of hits is HQ_{ji}, and the average accuracy is AAC_j. The calculation formula is as follows:

$$AC_{ji} = \frac{HQ_{ji}}{TQ_{ji}}$$

$$ACC_j = \frac{\sum_i HQ_{ji}}{\sum_i TQ_{ji}}$$

Table 6: Algorithm parameters corresponding to different models

Attack types	m	n	gamma	nu
Probe	0.001	8	1	0.1
DDOS	0.001	8	1	0.1
U2R	0.001	10	5	0.1
R2L	0.001	7	5	0.1

Through the comparison of experiments, it is found that when the parameter nu and gamma are larger, the fitting degree of OCSVM model is higher, and when the parameter nu and gamma are smaller, the fitting degree is lower. In the process of engineering implementation, increasing gamma parameter can improve the detection rate, but the false alarm rate of normal data will also increase. By reducing the nu parameter, the false alarm rate of normal data can be reduced, but the detection rate of abnormal attacks will also be reduced [1]. For example, in Table 10, R2L model has a low detection rate for guess_passwd attack, but when we choose to remove the 23, 24 witter sign, the detection rate for guess_passwd attack can reach 100%. This is because the contribution rate of 23 and 24 features is large when training the normal model, and the PCA algorithm still has a great impact on the data set features after the principal component extraction, but it can not show a good effect for guess_passwd attack detection. In fact, the removal of 23-dimensional and 24-dimensional features will also lead to a decrease in the detection rate of other types of attacks on R2L. This is because for other attacks on R2L, these two types of attacks can help the detection model to obtain a better distinction between abnormal data and normal data. In fact, the removal of 23-dimensional and 24-dimensional features will also lead to a decrease in the detection rate of other types of attacks against R2L. This is because for other subclass attacks in R2L attack type, these two features can help the detection model to obtain better differentiation between abnormal data and normal data. Therefore, it is found that for some special types of attacks, a special detection model can be established to improve the reliability of the intrusion detection system.

By comparing [1,8] with the detection model proposed in this paper, it can be found that the detection rate of the intrusion detection method proposed in this paper is similar to that of [1, 8] in the detection of Probe and DDOS attacks, but it is greatly improved in the detection of U2R and R2L attack types. The comparison results are shown in Figure 4. This shows that the Hierarchical PCA-OCSVM Model method proposed in this paper is better than OCSVM and SVM-ELM detection method in

Table 7: Detection results for different probe attacks

Probe Attack Types	Test Quantity	Accuracy Rate
Ipsweep	1247	93.5
Portsweep	1040	99.9
Nmap	232	88.8
Satan	1589	99.9
normal	5000	96.8

Table 8: Detection results for different DDOS attacks

DDOS Attack Types	Test Quantity	Accuracy Rate
teardrop	979	100
smurf	280790	99.9
pod	264	76.5
neptune	107201	99.9
normal	5000	98.5

Table 9: Detection results for different U2R attacks

U2R Attack Types	Test Quantity	Accuracy Rate
buffer_overflow	30	86.7
loadmodule	9	100
perl	3	100
rootkit	10	100
normal	5000	98.1

Table 10: Detection results for different R2L attacks

R2L Attack Types	Test Quantity	Accuracy Rate
warezclient	1020	93.5
warezmaster	20	100
multihop	7	85.7
imap	12	100
ftp_write	8	100
guess_passwd	53	30.2
normal	5000	98.7

Table 11: Detection results for different attacks

Attack types	Average Accuracy Rate
Probe	97.4
DDOS	99.9
U2R	92.3
R2L	90.6
normal	97.1

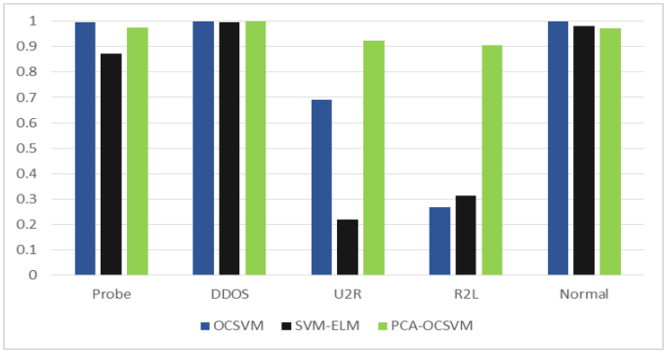

Figure 4: Comparison of hierarchical PCA-OCSVM detection accuracy with other methods

some functions, to some extent, it overcomes the problem of poor classification when OCSVM detects multiple attacks, and improves the effect of anomaly detection. The average accuracy for different attacks is shown in Table 11.

6 Conclusion

Firstly, this paper investigates the severe forms of current network security, and finds that the current network attacks can be divided into four major categories: Probe, DDOS, R2L and U2R. However, various kinds of small attack methods emerge in endlessly under different large categories, and traditional detection methods are gradually being broken down, so the intrusion detection method based on machine learning is becoming more and more important [3, 14].

According to the characteristics of OCSVM algorithm, this paper combines it with PCA algorithm to enlarge the difference between normal data and abnormal data. At the same time, the characteristics of KDDCUP99 data set are analyzed, and the data characteristics of different attack types are screened and preprocessed. The experiment of anomaly detection is carried out by using the OCSVM algorithm based on different kernel parameters. The disadvantages of OCSVM are improved by the multilayer anomaly detection model. By specifying the threshold value of the lowest characteristic contribution rate in PCA, the self-convergence function of PCA dimension parameters is realized, and the engineering implementation of PCA algorithm is optimized. Finally, based on KDDCUP99 data set, we test the accuracy of Probe, DDOS, R2L, U2R and different attack methods including these four attack types. The test results show that the detection accuracy of the proposed detection method for various types of attacks can reach 100%, and through the statistical average detection accuracy compared with previous studies, the proposed detection method is more excellent. In the follow-up research, we will focus on how to realize the automatic optimization and dynamic adjustment of parameters, so that the detection method can adapt

to different detection environment faster and get better detection effect. In the project implementation, the original KDDCUP99 data set is combined with the data set generated by the new attack [2, 9], so that the intrusion detection model can detect more kinds of network attacks.

References

[1] W. L. Al-Yaseen, Z. A. Othman, M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296–303, 2017.

[2] K. K. R. Amrita, "A hybrid intrusion detection system: Integrating hybrid feature selection approach with heterogeneous ensemble of intelligent classifiers," *International Journal of Network Security*, vol. 21, no. 3, pp. 438–450, 2019.

[3] K. K. R. Amrita, "Design of network threat detection and classification based on machine learning on cloud computing," *Cluster Computing*, vol. 22, no. 1, pp. 1–10, 2019.

[4] C. M. Bao, "Intrusion detection based on one-class SVM and SNMP MIB data," in *The Fifth International Conference on Information Assurance and Security*, pp. 346–349, Aug. 2009.

[5] M. Biba, L. Nishani, "Machine learning for intrusion detection in manet: A state-of-the-art survey," *Journal of Intelligent Information Systems*, vol. 46, no. 2, pp. 391–407, 2016.

[6] H. Duan, H. Hu, W. Qian, H. Ma, X. Wang, A. Zhou, "Incremental materialized view maintenance on distributed log-structured merge-tree," in *Pringer International Publishing AG, Part of Springer Nature*, pp. 682–700, 2018.

[7] L. Feng, Y. Wang, "Hybrid feature selection using component co-occurrence based feature relevance measurement," *Expert Systems with Applications*, vol. 102, pp. 83–99, 2018.

[8] J. Gong, M. Zhang, B. Xu, "An anomaly detection model based on one-class SVM to detect network intrusions," in *International Conference on Mobile Ad-hoc and Sensor Networks (MSN'16)*, pp. 102–107, 2016.

[9] A. Guezzaz, A. Asimi, Y. Asimi, Z. Tbatou, Y. Sadqi, "A global intrusion detection system using pcapsocks sniffer and multilayer perceptron classifier," *International Journal of Network Security*, vol. 21, no. 3, pp. 438–450, 2019.

[10] Z. Heng, *Design and Implementation of ELK based Network Security Log Management and Analysis System*, Beijing University of Posts and telecommunications, 2017.

[11] X. Jian, W. Xiaoqun, H. Zhihui, "Overview of china's internet security situation in 2018," *Confidential Science and Technology*, vol. 5, 2019.

[12] J. Jixue, H. Yingjie, Y. Zongmin, "Overview of machine learning application in intrusion detection," *Computer security*, no. 3, pp. 20–21, 2010.

[13] S. Khare, S. Y. Sait, A. Bhandari, "Multi-level anomaly detection: Relevance of big data analytics in networks," *Sadhana*, vol. 40, no. 6, pp. 1737–1767, 2015.

[14] Z. Liu, Y. Xin, L. Kong, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, no. 99, pp. 1–1, 2018.

[15] P. Padiya, U. Ravale, N. Marathe, "Feature selection based hybrid anomaly intrusion detection system using k-means and RBF kernel function," *Procedia Computer Science*, vol. 45, no. 39, pp. 428–435, 2015.

[16] Z. Qi, Z. Kun, "Application of machine learning in network intrusion detection," *Data Collection and Processing*, vol. 32, no. 3, pp. 479–488, 2017.

[17] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 7, pp. 1443–1471, 2001.

[18] D. Shin, D. Kim, Y. H. Kim, "Fast attack detection system using log analysis and attack tree generation," *Cluster Computing*, no. 2, pp. 1–9, 2018.

[19] J. Shouling, Q. Yaguan, L. Hongbo, "A toxic attack method for SVM based intrusion detection system," *Acta Electronica Sinica*, vol. 47, no. 1, pp. 59–65, 2019.

[20] M. Touahria, S. Maza, "Feature selection for intrusion detection using new multi-objective estimation of distribution algorithms," *Applied Intelligence*, vol. 49, no. 12, pp. 4237-4257, 2019.

[21] University of California, "Kdd cup 1999 data," *The UCI KDD Archive Information and Computer Science*, 2019. (`http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`)

[22] H. Xiangjing, H. Liang, C. Zemao, "Research on intrusion detection algorithm based on improved k-means clustering," *Computer and digital engineering*, vol. 6, pp. 1145–1149, 2017.

[23] L. Xing, *Research and Design of DoS Attack Detection System based on Snort*, Beijing University of Posts and telecommunications, 2015.

[24] W. Xu, Y. Xiao, H. Wang, "Parameter selection of gaussian kernel for one-class SVM," *IEEE Transactions on Cybernetics*, vol. 5, pp. 927–939, 2015.

[25] Z. Yan, Z. Jin, Y. Cui, "Survey of intrusion detection methods based on data mining algorithms," in *Proceedings of International Conference on Big Data Engineering*, pp. 98–106, June 2019.

[26] W. Zhao, Q. Liu, P. Li, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, no. 99, pp. 12103–12117, 2018.

[27] S. Zhonglin, N. Lei, "Pca-akm algorithm and its application in intrusion detection," *Computer Science*, no. 2, pp. 41, 2018.

Biography

Yanpeng Cui, born in 1978, female, doctoral student, lecturer. The main research fields are electronic warfare signal processing, machine learning, intrusion detection and so on..

Zichuan Jin, born in 1995, male, master's degree student, mainly studies machine learning, network operation and maintenance, intrusion detection and other directions.

Jianwei Hu, born in 1973, male, Ph. D. doctoral student. He mainly studies computer network, industrial control system, network hardware and software security and attack-defense confrontation.