Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3056

NETWORK INTRUSION DETECTION BASED ON RANDOM FOREST ALGORITHM

RongPeng Yan
Sports Engineering College, Beijing Sport University, Beijing 100091, China.
Corresponding Email: 18840697359@163.com

Abstract: With the rapid development of Internet technology, network security problems are becoming increasingly serious, and the frequency and complexity of network attacks are increasing, posing a serious threat to personal privacy, corporate interests and even national security. For the problem of redundant feature interference and dimensional disaster in high-dimensional network traffic data, this paper compares the effectiveness of feature screening and dimensionality reduction techniques, such as ANOVA, chi-square test and PCA, respectively, for the removal of irrelevant features in high-dimensional network traffic data, and the experimental results show that PCA solves the problem of high complexity of high-dimensional data processing and effectively improves the classification performance and operational efficiency of the model. Therefore, this study innovatively proposes a hybrid intrusion detection model that integrates Principal Component Analysis (PCA) and Random Forest (RF), and adopts a grid search algorithm to automate the optimization of the hyper-parameter set of the Random Forest, and finally the model has an accuracy of 99.81% in the test set, which indicates that it performs well in classifying the attack and normal traffic. Overall, the model provides an efficient and accurate solution for network intrusion detection, which has important reference value for future research and practical applications.
Keywords: Principal component analysis; Random forest; Network intrusion detection; Feature selection

1 INTRODUCTION

With the rapid development of Internet technology, the network has been deeply integrated into all aspects of people's lives, greatly facilitating daily life and promoting social progress and economic development. However, network security issues are becoming increasingly serious, a variety of known or unknown network attacks occur frequently, posing a serious threat to personal privacy, corporate interests and even national security. In this context, Intrusion Detection System (IDS), as an important line of defense for network security, has become more and more prominent, and IDS, by collecting key information in the network, can detect and warn of abnormal behaviors and intrusion attacks in the network in a timely manner, providing a strong guarantee for network security. Traditional IDS from the initial use of audit data to track the user's suspicious behavior, to propose the first real-time network intrusion detection expert system model, and then use the state transition analysis to improve the model, the identification of intrusion attacks has always been a hot issue in the field of network security[1-4].

In recent years, many scholars have devoted themselves to introducing machine learning techniques into the field of intrusion detection to improve the accuracy and efficiency of detection. For example, Jintai Wei et al. proposed an intrusion detection system based on information gain and random forest classifier, which utilizes a synthetic minority oversampling algorithm to solve the data imbalance problem and feature selection by information gain to effectively improve the minority class anomaly detection rate. Zhou Ying et al. used PCA and KNN algorithm for feature selection, and experiments show that KNN algorithm performs well on small datasets and can improve the accuracy and reduce the false alarm rate of intrusion detection system[5]. On the other hand, Zhu Linjie et al. proposed an intrusion detection method based on the combination of mutual information feature selection and KNN classifier, which improves the accuracy of intrusion detection by simplifying the model and optimizing the variable selection[6]. However, there are some shortcomings in machine learning methods, for example, some of them have high computational complexity when dealing with high-dimensional data, which leads to low operational efficiency and makes it difficult to meet the demand of real-time detection. For example, although the feature selection methods based on complex optimization algorithms such as particle swarm optimization (PSO) have improved the feature selection effect, the running time on large-scale datasets is too long, which restricts its practical application[7]. On the other hand, some studies have paid insufficient attention to the differences in the detection performance of different attack types in datasets. Different attack types tend to exhibit category imbalance in the datasets, and some of the existing methods have low detection accuracy for a few categories, which cannot comprehensively and effectively cope with diverse cyber-attacks. In addition, some studies ignore the correlation between features in the feature selection process, which may lead to redundant information in the selected features, affecting the performance and efficiency of the model. For example, the simple filtered feature selection method selects features only based on their own statistical characteristics and fails to fully consider the interactions between features, which affects the optimality of the feature subset[8].

Therefore, in order to remove irrelevant features to improve the classification performance of the model, and at the same time consider different models to achieve the optimal effect, this paper proposes a PCA-based detection model for intrusion detection in random forest networks[9-10]. The main contributions contain points:

1) The original datasets is analyzed and preprocessed using methods such as solo thermal coding and binary classification.
2) Compare the effect of three basic machine learning models, logistic regression, support vector machine and random forest, from the perspective of machine learning without feature screening.
3) Using the random forest model, combined with ANOVA, chi-square test for feature screening, and finally using PCA for data dimensionality reduction.

2 ALGORITHMIC PRINCIPLE

In this section, the overall framework of PCA-based random forest intrusion detection model is firstly given, followed by the specifics of feature selection and classification models.

2.1 Feature Screening

2.1.1 ANOVA (analysis of variance)

Based on the variance decomposition of the observed variables, the total variance is divided into between-group and within-group variance. Intergroup variance is caused by systematic differences due to different levels of the factors, and intragroup variance is caused by random factors such as sampling error. By comparing the between-group variance and within-group variance, it is determined whether the effect of the factors on the observed variables is significant or not. The steps of the algorithm are as follows.
1. Calculate the sum of squares (SST): the sum of the squared deviations of all data points from the overall mean, reflecting the total variation in the data.

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2 \tag{1}$$

where k is the number of groups, n_i is the sample size of group i, X_{ij} is the j observation of group i, and $\overline{X}$ is the total mean.
2. Calculate the sum of squares between groups (SSA): the sum of squares of the deviations of the group means from the total mean, reflecting the between-group variation.

$$SSA = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2 \tag{2}$$

where $\overline{X}_i$ is the mean of group i.
3. Calculate the sum of squares within groups (SSE): the sum of squares of the deviations of the data within each group from the group mean, reflecting the within-group variation.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2 \tag{3}$$

4. Calculate the mean square: the between-group mean square (MSA) is the between-group sum of squares divided by the between-group degrees of freedom, and the within-group mean square (MSE) is the within-group sum of squares divided by the within-group degrees of freedom.

$$MSA = \frac{SSA}{k-1} \tag{4}$$

$$MSE = \frac{SSE}{N-k} \tag{5}$$

where N is the total sample size.
5. Calculate the F statistic: the F value is equal to the between-group mean square divided by the within-group mean square, and is used to test whether the group means are equal.

$$F = \frac{MSA}{MSE} \tag{6}$$

6. Determination of the level of significance and critical value: Usually, the level of significance α (e.g., 0.05) is chosen, and the corresponding critical value is found according to the F distribution table. If the F-statistic is greater than the critical value, the original hypothesis is rejected and the group means are considered not all equal; otherwise, the original hypothesis is not rejected.

2.1.2 Chi-square (math.) test

It is used to test the correlation between categorical variables (independence test) or to test whether the actual distribution of categorical variables matches the theoretical distribution (goodness-of-fit test). The basic idea is to compare the difference between the actual and theoretical frequencies, and to determine whether this difference is caused by random factors through the chi-square statistic. The steps of the algorithm are as follows.
1. List the actual and theoretical frequencies: the actual frequency is the observed data, and the theoretical frequency is the expected frequency calculated according to the hypothesis.

$$E_{ij} = \frac{n_i \cdot n_j}{N} \tag{7}$$

Where E_{ij} is the theoretical frequency of the j column of the i row, n_i is the marginal total of the i row, n_j is the marginal total of the j column, and N is the total sample size.
2. Calculate the chi-square statistic: the chi-square value is equal to the square of the difference between the actual frequency and the theoretical frequency of each cell divided by the theoretical frequency, and then the results of all cells are added together.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (8)$$

Where O_{ij} is the actual frequency, r is the number of rows and c is the number of columns.

3. Determine the degree of freedom and significance level: the degree of freedom is $(r-1)\times(c-1)$, and the significance level is usually chosen as 0.05.

4. Check the chi-square distribution table to get the p-value: according to the calculated chi-square statistic and degrees of freedom, check the chi-square distribution table to get the corresponding p-value. If the p-value is less than the significance level, the original hypothesis is rejected and the categorical variables are considered to be related; otherwise, the original hypothesis is not rejected.

2.1.3 Principal component analysis (PCA)

A commonly used dimensionality reduction technique, which projects the data into a low dimensional space by linear transformation while retaining the main features and information of the data as much as possible, the following is the principle of the algorithm,.

1. Data standardization: each eigenvalue of the data is transformed into a distribution with a mean of 0 and a variance of 1 to eliminate the effects of different feature measures and measure sizes.

$$X = \frac{X-\mu}{\sigma} \quad (9)$$

where X is the original data, μ is the mean of the data and σ is the standard deviation of the data.

2. Calculate the covariance matrix: the covariance matrix is used to measure the correlation between different features, and its diagonal elements indicate the variance of each feature, and the off-diagonal elements indicate the covariance between different features.

$$C = \frac{1}{n-1}X^T X \quad (10)$$

where X is the standardized data matrix and n is the number of samples.

3. Calculate the eigenvalues and eigenvectors of the covariance matrix: the eigenvalues indicate the degree of data dispersion in the direction of the corresponding eigenvectors, and the larger the eigenvalues are, the greater the change in the data in the direction, and the more information they contain.

$$Cv_i = \lambda_i v_i \quad (11)$$

where λ_i is the eigenvalue and v_i is the corresponding eigenvector.

4. Select principal components: arrange the corresponding eigenvectors in the order of the eigenvalues from the largest to the smallest, and select the first k eigenvectors as the principal components, and the direction of these principal components is the direction with the largest variance in the data.

5. Data projection: Project the original data into the low-dimensional space composed of principal components to get the dimensionality reduced data.

$$Y = XW \quad (12)$$

where W is the matrix consisting of the first k eigenvectors and Y is the dimensionality reduced data.

2.2 Model Screening

2.2.1 Random forest

An integrated learning method that performs classification or regression by constructing multiple decision trees and integrating their results. The core idea is to use randomness and diversity to reduce the overfitting risk of a single decision tree and improve the generalization ability and stability of the model, the steps of the algorithm are as follows,.

1. Data sampling

Generate multiple sub-datasets from the original datasets by sampling with put-back (Bootstrap sampling).

2. Construct decision tree

For each sub-data set, construct a decision tree.

In the process of constructing a decision tree, each time the best splitting feature is selected, a subset of all features is randomly selected, and then the best splitting feature is searched for in that subset.

3. Voting or averaging

Classification problem: the final classification result is determined by the voting results of all decision trees, i.e., the category with the highest number of occurrences is selected.

Regression problem: determine the final regression result by averaging the predictions of all decision trees.

4. Prediction

Use the trained Random Forest model to predict the new data and get the final classification or regression result.

3 CASE STUDY

3.1 Experimental Environment

This paper uses python language for model construction. The experimental environment is Intel64 Family 6 Model 154 step3, GenuineIntel, 15.69GB RAM, Windows-11-10.0.22631-SP0 64-bit operating system.

3.2 Data Set

In this paper, the NSL-KDD datasets is selected for experiments, the datasets contains 43 features, based on the analysis of these feature data, the category of the attack (normal vs. attacked) can be finally identified.

3.3 Data Preprocessing

Firstly, missing value detection is performed on the data set and it is found that there are no missing values. Then outlier detection is performed on the datasets and it is found that there are data outliers in the original datasets, as the presence of outliers may have an impact on the model results, the results of outliers using the quartile method are shown in Figure 1 below.

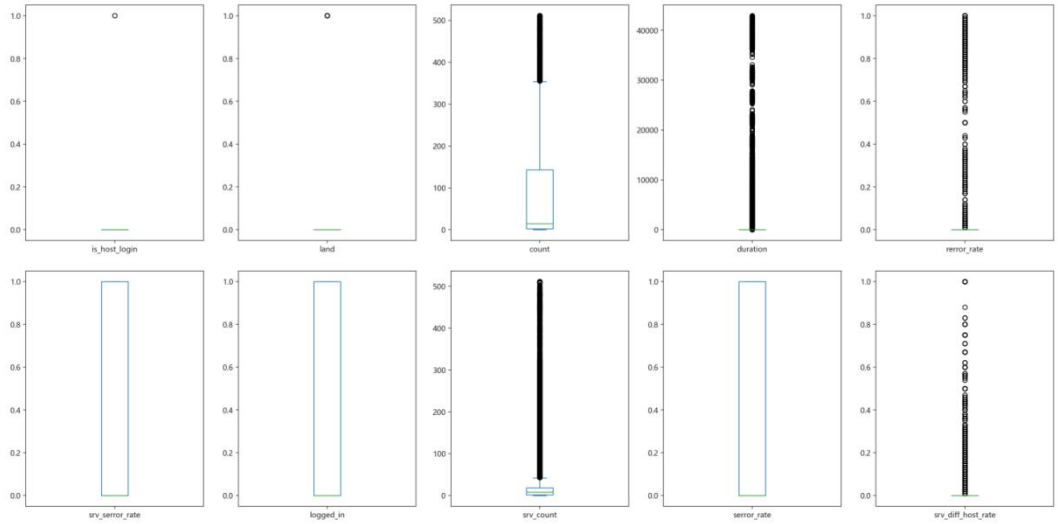

Figure 1 Data Set Outlier Plots

Observing the above figure 1, we can see the distribution of outliers of different features, in which there are more outliers in features such as duration, srv_count, etc, while there are less outliers in features such as 'wrong_fragment'. After the quartile method processing, this paper plots the distribution of raw data as shown in Figure 2 below.

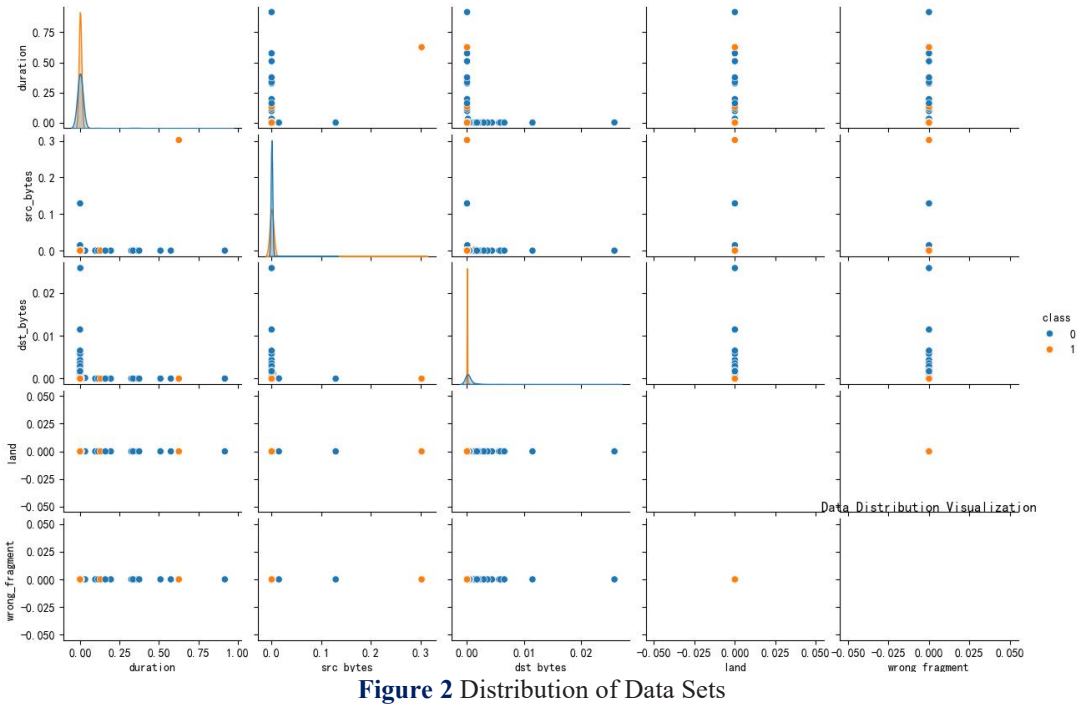

Figure 2 Distribution of Data Sets

Through the above Figure 2, the difference in the distribution of different features under the two classes is found. Among them, the distribution of duration is more similar under class 0 and class 1, but the samples of class 1 may be longer in duration; while the distribution of src_bytes is more different under class 0 and class 1, the distribution of

class 0 is more centralized, while the distribution of class 1 is more dispersed, and some of the samples have higher src_ bytes value is higher.
Next, the type of attacked was changed to both attacked and normal using binary classification, as shown in Figure 3 below.

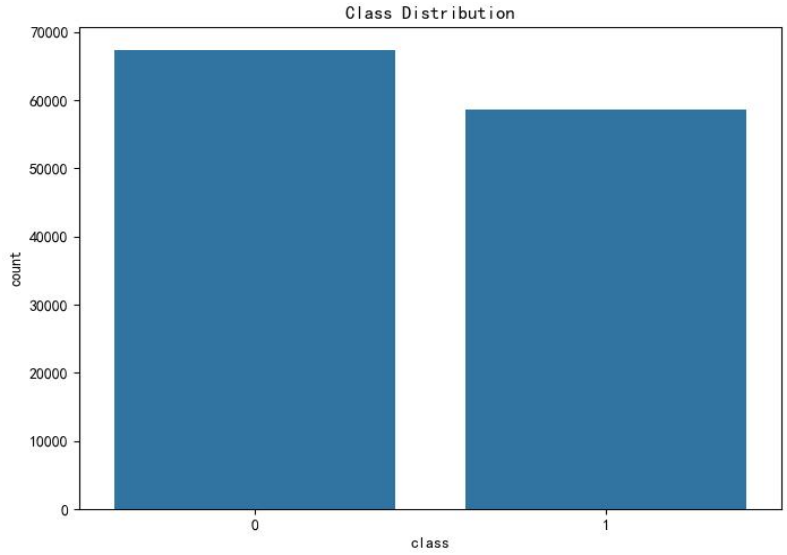

Figure 3 Diagram of the Results of the Binary Classification

In Figure 3, it can be seen that the number of samples in class 0 (normal) is slightly more than the number of samples in class 1 (attack), but both are generally more evenly distributed.
Finally, the data is normalized and uniquely hot coded to facilitate follow-up.

3.4 Feature Selection

First of all, based on the observation of the features, it can be found that the data is divided into two kinds of numerical features and non-numerical features, and for the numerical features ANOVA is used for processing.
In the ANOVA results, it can be seen that features such as srv_rerror_rate, dst_host_rerror_rate, and dst_host_srv_serror_rate have high F-values, which indicates that there are significant differences between these features in different categories. Whereas, features such as duration, num_root, and num_file_creations have low F-values, indicating that they are not significantly different between categories.
Then, the chi-square test is performed on the non-numeric features and the results are shown in Figure 4 below.

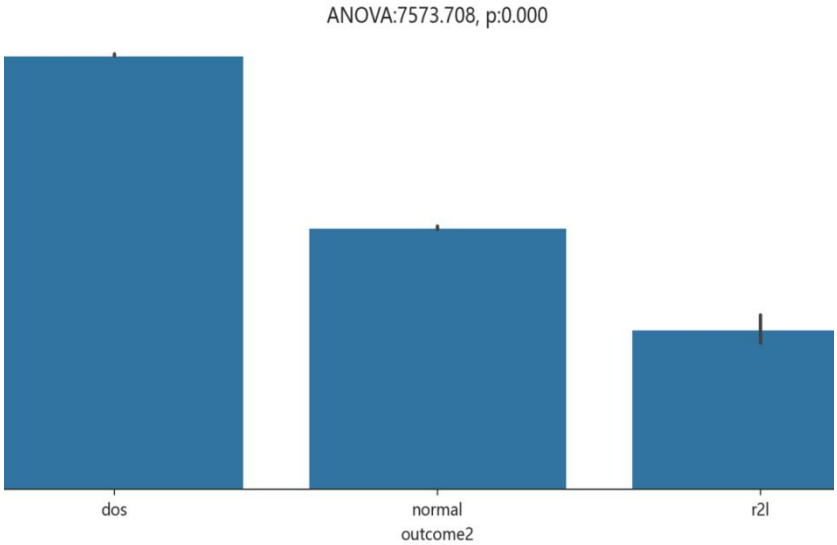

Figure 4 Chart of Chi-Square Test Results

From Figure 4, the chi-square test scores for the non-numeric features and the target variable can be obtained. Among them, class_label_attack and class_label_normal have the highest scores, indicating that these two categories are

significantly different in the datasets. Other features such as 'flag_S0', 'flag_SF', and 'service_http' also have high scores, indicating that they are strongly correlated with the target variable.
Finally, after the results of ANOVA and chi-square test, these features were subjected to PCA dimensionality reduction, in which the number of selected principal components and the results of the model are shown in Figure 5 below.

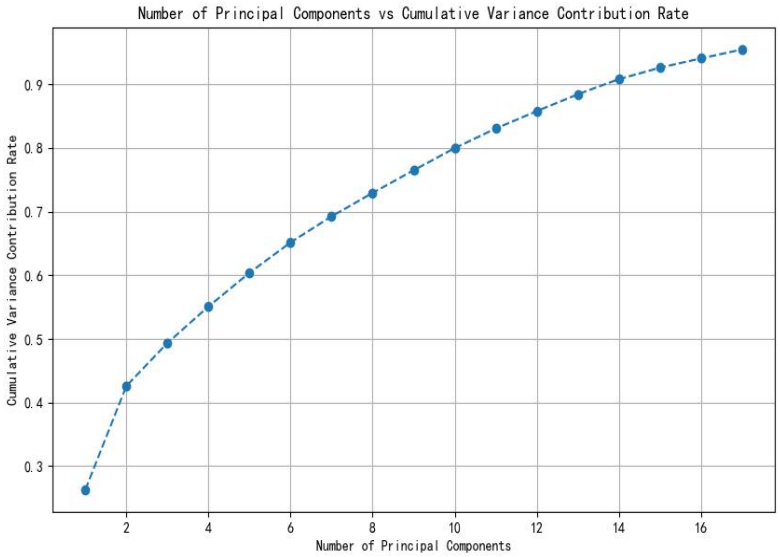

Figure 5 Principal Component Selection Chart

From Figure 5, it can be found that the cumulative variance contribution rate increases gradually with the increase of the number of principal components, and at 17 principal components, the cumulative variance contribution rate is more than 95%, which indicates that these principal components can better retain the variance information of the original data, so, by PCA dimensionality reduction, 17 principal components are retained, which can better retain the variance information of the original data, and can distinguish between samples of different categories.

3.5 Model Building

First, the datasets is classified test set 20%, training set 80%, and set the random seed to 42 to ensure the reproducibility of the results, next choose the three most basic machine learning models: logistic regression, support vector machine and random forest, and analyze the results by constructing the model to choose the appropriate model, as shown in Table 1-4 and Figure 6 below.

Table 1 Logistic Regression Model Indicators

Logistic Regession	Precision	Recall	F1-Score	Support
Attack	0.95	0.97	0.96	11773
Normal	0.97	0.96	0.97	13422
Macro Avg	0.96	0.97	0.96	25195
Weighted Avg	0.96	0.96	0.96	25195

Table 2 Support Vector Machine Model Metrics

Support Vector Machine	Precision	Recall	F1-Score	Support
Attack	0.98	0.99	0.99	11773
Normal	0.99	0.99	0.99	13422
Macro Avg	0.99	0.99	0.99	25195
Weighted Avg	0.99	0.99	0.99	25195

Table 3 Random Forest Model Indicators

Random Forest	Precision	Recall	F1-Score	Support
Attack	1.00	1.00	1.00	11773
Normal	1.00	1.00	1.00	13422
Macro Avg	1.00	1.00	1.00	25195
Weighted Avg	1.00	1.00	1.00	25195

Table 4 Comparison of Results from Different Models

Methodologies	Accuracy
Logistic Regression	0.9646

Support Vector Machine	0.9894
Random Forest	0.9981

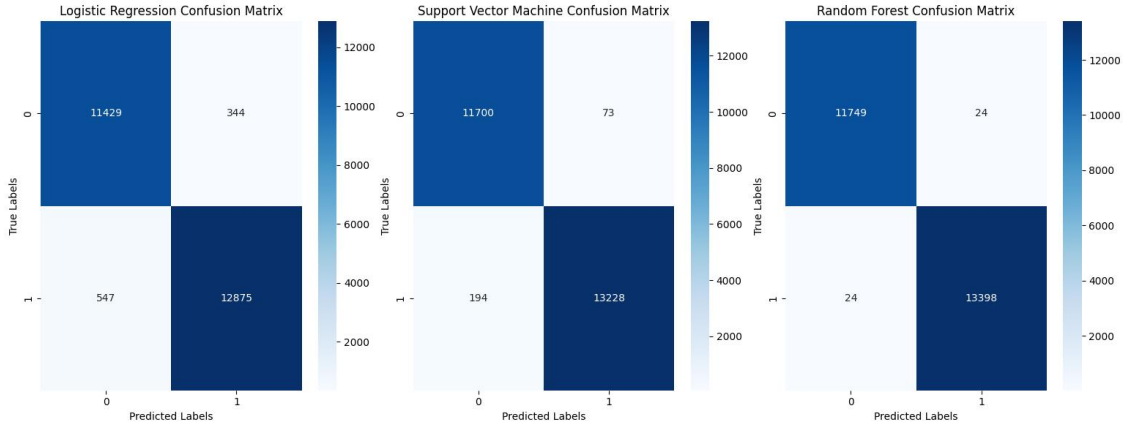

Figure 6 Plot of Results of Comparison of Confusion Matrices of Different Models

Combining the above Tables 1-4 and Figure 6, among the three models, Logistic Regression, Support Vector Machine and Random Forest, the Random Forest model achieves the best performance, with an accuracy of 0.9981, and all the metrics in the classification report are 1.00, which indicates that it classifies the attack and normal traffic on the training set almost perfectly.The Support Vector The Support Vector Machine model, with an accuracy of 0.9894 and all the indexes in the classification report around 0.99, performs well and stably, while the Logistic Regression model, with an accuracy of 0.9646, is relatively low among the three models, and its confusion matrix shows that there are more misclassification cases. On balance, Random Forest has a significant advantage in this task, so Random Forest is chosen as the main model.

3.6 Model Evaluation

For parameter optimization of the random forest model selected above, the choice was made to use the grid search method as shown in Table 5 below.

Table 5 Optimal Parameter

	n_estimators	min_samples_split	max_depth	min_samples_leaf	max_features
optimal parameter	100	2	None	1	sqrt

As can be seen in Table 5, the best parameters identified by the grid search are 'n_estimators=100', 'max_depth=None', 'min_samples_split=2', 'min_samples_leaf=1', and 'max_features='sqrt''. The random forest model using these parameters achieved a high accuracy of 99.81% on the test set, indicating that the model performs well in distinguishing between attack and normal traffic. The classification report shows that the model achieves 1.00 in precision, recall, and F1 scores for both attack and normal categories, indicating that the model classifies samples of both categories almost perfectly on the training set.

4 CONCLUSION

In this paper, a PCA-based random forest network intrusion detection model is proposed to cope with the complex challenges in the field of network security. Firstly, the NSL-KDD datasets is meticulously preprocessed, including outlier detection and processing, binary conversion; ANOVA, chi-square test and PCA techniques are applied for feature screening and dimensionality reduction, which effectively removes irrelevant features and reduces the dimensionality of the data; three models, namely, logistic regression, support vector machine, and random forest, are systematically compared, and random forest is finally selected as the optimal model and its parameters are optimized by lattice Random Forest is finally selected as the optimal model, and its parameters are optimized by the grid search method. The experimental result is that the random forest model achieves a high accuracy of 99.81% on the test set, and all the indexes of the classification report are 100%, which indicates that it performs well in the classification of attack and normal traffic, and can provide effective and accurate classification effect for network intrusion detection.

Although the PCA-based random forest network intrusion detection model proposed in this paper achieved a high accuracy rate (99.81%) in the experiment, there are still many shortcomings. The feature engineering complexity is high, and the PCA dimension reduction may lose some information; the random forest model has limited computational efficiency when dealing with large-scale high-dimensional data, which makes it difficult to meet the real-time detection demand, the adaptability to new unknown attacks is insufficient, and the imbalance between the number of attack samples and normal samples in the datasets may affect the detection performance. Future research further explores deep learning models (CNN, RNN, or Transformer) to automatically learn feature representations, reduce the dependence on

manual feature engineering, and improve the classification performance and real-time performance of the model, as well as combining methods such as data augmentation techniques (SMOTE) and generative adversarial networks (GAN) to enhance the ability to deal with unknown attacks and data imbalance problems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Liang Junwei, Yang Geng, Ma Maode, et al. Collaborative intrusion detection system for industrial CPS based on secure federated distillation GAN. Journal of Communications, 2023, 44(12): 230-244.
[2] Zhao Y. Practice of computer network security technology in network security maintenance. China New Communication, 2023, 25(20): 101-103.
[3] He Feifei. Research and implementation of real-time network intrusion detection method based on deep learning. Ningxia University, 2022.
[4] Li B. Strict minimal beacon solving and net state analysis based on Petri net decomposition technique. Xi'an University of Electronic Science and Technology, 2021.
[5] Zhou Y. PCA and KNN feature selection for network intrusion detection. Computer and Network, 2021, 47(22): 48-49.
[6] Zhu L J, Zhao G P, Kang L H. Intrusion detection method based on the combination of MI feature selection and KNN classifier. Gansu Science and Technology, 2022, 38(15): 33-36.
[7] Jianhua Chen, Zhangqian Wu, Wei Song. A particle swarm optimization algorithm incorporating precocity detection mechanism and opposing stochastic wandering strategy. Computer Applications, 2024, 44(S2): 123-128.
[8] Hong Zhou,Yang Gang, Yang Jinsong, et al. Feature selection of tag library with joint filtered and embedded samples. Electronic Design Engineering, 2024, 32(22): 146-150.
[9] Meng Yuru, Ren Xiaoling, Wei Ziyi. A guided filter image fusion method based on the combination of PCA and NSCT. Computer and Digital Engineering, 2024, 52(10): 3116-3120.
[10] Guo Yu-Han, Zhu Ru-Shi. A dynamic point-of-entry recommendation algorithm based on multimodal deep forest and iterative Kuhn-Munkres. Computer Application Research, 2024, 41(12): 3634-3644.