An Effective Intrusion Detection Model based on Random Forest Algorithm with I-SMOTE

Weijinxia[1], Longchun[1,2], Wanwei[1,2], Zhaojing[1], Duguanyao[1,2] and Yangfan[1]
[1]*Department of Security, Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China*
[2]*Chinese Academy of Sciences University, Beijing, 101408, China*

Keywords: Intrusion Detection, Class Imbalance, I-SMOTE, Feature Reduction, Random Forest.

Abstract: With the wide applications of network in our daily lives, network security is becoming increasing prominent. Intrusion detection systems have been widely used to detect various types of malicious network which cannot be detected by a conventional firewall. Therefore, various machine-learning techniques have been proposed to improve the performance of intrusion detection system. However, the balance of different data classes is critical and will affect detection performance. In order to reduce the impact of class imbalance of the intrusion dataset, this paper proposes a scheme that applies the improved synthetic minority oversampling technique (I-SMOTE) to balance the dataset, employs correlation analysis and random forest to reduce features and uses the random forest algorithm to train the classifier for detection. The experimental results based on the NSL-KDD dataset show that it achieves a better and more robust performance in terms of accuracy, detection rate, false alarms and training speed.

1 INTRODUCTION

The aim of cloud computing is to provide on-demand network security services, which are realized over the Internet. Therefore, the security and privacy of network data are crucial issues in cloud computing. With the development of network technology and the continuous expansion of cloud computing applications, the network attacks have become more diverse (Amin Hatef M et al., 2018). Also, people are facing to security threats from the Internet. Therefore, to protect the security of computer and Internet, the various machine-learning techniques have been proposed to improve the performance of intrusion detection systems. However, with the complexity of intrusion or attack techniques, some existing approaches exhibit limited ability. Thus, the performance optimization of the intrusion detection system has received much attention (Luo et al., 2014) (Tan et al., 2019).

Much related work focuses on the task of intrusion detection based on various artificial intelligence and machine learning algorithms. There exist many recent studies, which combines or assemble several algorithms in order to improve detection model's performance, such as detection rate, accuracy, area under curve and false alarm rates.

However, there are two limitations to existing works. First, although advanced and complicated detection models have been built, very few have considered the effects of high dimensionality and sample imbalance simultaneously, which are important problems in improving model's performance. Second, the time taken for training and testing model indicates that their models are difficult to apply in actual situations.

Therefore, in this work, we propose a multi-classification scheme that applies the improved synthetic minority oversampling technique (I-SMOTE) to balance the dataset, employs correlation analysis and random forest to reduce features and uses the random forest algorithm to train the classifier for detection. Details are as follows:

1) The I-SMOTE is an improved oversampling algorithm and is used to solve the imbalance problem of dataset, which can consider impacts of majority samples and minority samples on oversampling results simultaneously.
2) Correlation analysis and random forest are applied to reduce the feature dimensions. The purpose of correlation analysis is to calculate correlation between each feature and classification label, and form a new feature sequence according to order of correlation values from big to small.

Weijinxia, ., Longchun, ., Wanwei, ., Zhaojing, ., Duguanyao, . and Yangfan, .
An Effective Intrusion Detection Model based on Random Forest Algorithm with I-SMOTE.
DOI: 10.5220/0010393801750182
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1*, pages 175-182
ISBN: 978-989-758-509-8

Then, the ACC of random forest model is regarded as fitness of feature reduction, and the dimension corresponding to the largest ACC is the optimal dimension.

3) Considering the random forest is better than other commonly used detection models in training and testing time, therefore, the random forest is used as the multi-attack detection classifier in this work.

2 RELATED WORKS

Many machine learning approaches have been applied to construct intrusion detection systems. The most used supervised algorithms include decision trees, support vector machine and K-nearest neighbors, random forest (Horng et al., 2011, Manjula et al., 2016).

(Bamakan et al., 2016) introduced an SVM-based intrusion detection model. In their paper, the time-varying chaos particle swarm optimization was used to optimizing parameters of SVM classifier. An SVM-based on intrusion detection approach that combined the genetic algorithm (GA) and kernel principal component analysis (KPCA) was proposed by (Kuang et al., 2014). The KPCA was used to reduce feature dimensions and the GA was applied to optimize parameters of SVM. (Horng et al., 2011) proposed an SVM-based intrusion integrated with the BIRCH hierarchical clustering algorithm which was used to preprocess the original data. Thus, the data used for training intrusion detection model has a high quality.

Generally speaking, the intrusion detection system should produce low false-positive rate and process imbalanced datasets, a large number of features and a large amount of training and testing data (Chen Tongbao et al., 2020). Based on this scenario, random forest algorithm presents great advantages in training time and prediction time compared with SVM (Li et al., 2020; Chauhan et al., 2013; Amira et al., 2017) (Manjula et al., 2016; Golrang et al., 2020). In Chauhan et al.'s paper (Chauhan et al., 2013), top-ten classification algorithms namely J48, Logistic, IBK, JBayesNet, SGD, PART, Rip, Random Forest, Random Tree and REPTree were selected for experimental comparison. In experimental results, the Random Forest classifier with 99.75% accuracy has got the first position in ranking. Also, sensitivity of Random Forest classifier was also highest compared to others. In (Manjula et al., 2016), classification and predictive models for intrusion detection were proposed by using machine learning classification namely Random Forest, Gaussian Naïve Bayes, Support Vector Machine and Logistic Regression. These models were experimented with NSL-KDD dataset. The experimental results illustrated that Random Forest Classifier showed great advantages than other methods in identifying whether the data is normal or abnormal. To improve classification accuracy and reduce training time, in (Li et al., 2020) proposed an effective deep learning method, namely AE-IDS (Auto-Encoder Intrusion Detection System) based on random forest algorithm. This method constructed the training set with feature selection and feature grouping by using random forest. The experimental results show that the proposed method is superior to traditional machine learning based intrusion detection methods in terms of easy training, strong adaptability, and high detection accuracy.

In the above works, the random forests were used as a separate classification model. In recent years, there also exist many hybrid models that use Random Forest. Malik et al. (Malik et al., 2011; (Malik et al., 2013; Malik et al., 2015) applied Particle Swarm Optimization and Symmetrical Uncertainty for feature selection and Random Forest model as the classifier to detect probe attacks. A hybrid model based Random Forest and Neural Networks was introduce by (Zhong et al., 2011), where the Neural Networks algorithm is used to classify and the random forest is applied to reduce feature dimension. (Hasan et al., 2014) proposed a hybrid model using support vector machine and random forest. The experimental results showed that detection performance of the hybrid model was better than the individual models.

Considering the imbalance of dataset, (Tesfahun et al., 2013) used SMOTE to reduce the imbalance and Random Forest algorithm as a classifier to establish intrusion detection model. (Tan et al., 2019) proposed an intrusion detection algorithm based on SMOTE and random forest. The simulations are conducted on a benchmark intrusion dataset, and the accuracy of model has reached 92.39%. However, only K-neighbors operations in minority class were considered for oversampling operations in SMOTE, and did not consider the majority class samples appeared in K-neighbors operations, which easily led to the invariance of minority class density after oversampling (Wang et al., 2018).

Based on the above analysis, we propose a scheme that applies the improved synthetic minority oversampling technique (I-SMOTE) to balance the dataset, employs correlation analysis and random forest to reduce features and uses the random forest algorithm to train the classifier for detection. First, the I-SMOTE is an improved oversampling algorithm and

is used to solve the imbalance problem of dataset, which can consider impacts of majority samples and minority samples on oversampling results simultaneously. Second, correlation analysis and random forest are applied to reduce the feature dimensions. The purpose of correlation analysis is to calculate correlation between each feature and classification label, and form a new feature sequence according to order of correlation values from big to small. Then, the ACC of random forest model is regarded as fitness of feature reduction, and the dimension corresponding to the largest ACC is the optimal dimension. Finally, considering the random forest is better than other commonly used detection models in training and testing time, therefore, the random forest is used as the classifier in this work.

3 PRELIMINARY

3.1 Correlation Analysis

Correlation metrics are widely applied in machine learning and statistical correlation analysis to evaluate the correlation between features. The selection of correlation metrics affects the efficiency of feature selection greatly. The correlation degree between two random variables is usually measured by entropy and mutual information which are defined in information theory.

Definition 1. For a discrete feature vector $X \in \{x_1', x_2', \cdots, x_n'\}^T$, its probability distribution can be expressed as $\{p(x_1'), p(x_2'), \cdots, p(x_n')\}$, the entropy of feature X is as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i') \log_2 p(x_i') \tag{1}$$

If all the values of X are the same, then the entropy of X is 0. Thus, the feature X is useless for data classification.

Definition 2. For two discrete features $X \in \{x_1', x_2', \cdots, x_n'\}^T$ and $Y \in \{y_1, y_2, \cdots, y_m\}$, their joint probability density is $p(x_i', y_j), 1 \le i \le n, 1 \le j \le m$, and conditional density is $p(x_i' \mid y_j)$, then entropy of X under the condition Y can be expressed as

$$H(X \mid Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i', y_j) \log_2 \frac{p(y_j)}{p(x_i', y_j)} \tag{2}$$

The mutual information is generated and derived from entropy. For two features X and Y in one dataset, the mutual information between them is as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i', y_j) \log_2 \frac{p(x_i', y_j)}{p(x_i') p(y_j)} \end{aligned} \tag{3}$$

The mutual information has the following characteristics:

Symmetry: $I(X;Y) = I(Y;X)$

Monotonic: if $A \subseteq B \subseteq C$, then $I(A;C) = I(B;C)$

The mutual information reflects the amount of information shared between two random variables. The greater value of the mutual information, the greater correlation between the two variables. If the mutual information between two variables is 0, the two variables are completely uncorrelated and statistically independent in probability.

3.2 SMOTE Algorithm (Wang et al., 2018)

The SMOTE algorithm oversamples minority samples. For a certain minority sample x , find the K minority samples closest to it. If up-sampling ratio is N , we randomly select N samples from K minority samples, $y_1', y_2', \cdots, y_N'$. Then N new minority samples $X_{new1}, X_{new2}, \cdots, X_{newN}$ can be generated by the following equation:

$$X_{newj} = x + rand(0, 1) \cdot (y_j' - x), j = 1, 2, \cdots, N \tag{4}$$

Where $rand(0, 1)$ is a random number in $(0, 1)$.

4 PROPOSED FRAMEWORK FOR INTRUSION DETECTION

In this section, we list the main steps of our paper. First, we introduce an improved SMOTE (I-SMOTE) algorithm. Second, correlation analysis is applied to arrange features of dataset after oversampling. The purpose of correlation analysis is to calculate correlation between each feature and classification label, and form a new feature sequence according to order of correlation values from big to small. Then, the ACC of random forest model is regarded as fitness of feature reduction, and the dimension corresponding to

the largest ACC is the optimal dimension. Finally, the dataset with feature reduction is used to train the random forest classifier. The specific procedures of the proposed intrusion detection are illustrated in Fig.1. As shown in Fig.1, the frame work of the detection model mainly consists of three parts: Oversampling based on I-SMOTE, feature reduction and intrusion detection.

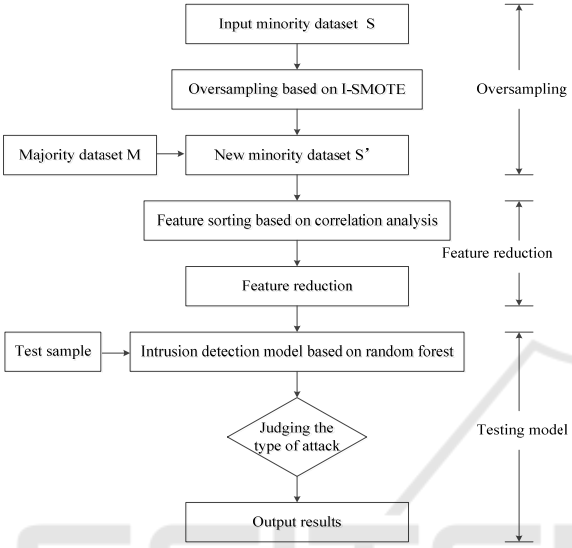

Figure 1: The procedures of the proposed intrusion detection.

4.1 Dataset oversampling based on I-SMOTE

For each sample x in minority samples S, we search for k samples closest to x, where k is obtained by cross-validation. In k samples, if the number of samples of minority class is larger than majority class, x is a safe sample; if the number of samples of minority class is less than majority class and there exist minority class samples, x is a dangerous samples; if all k samples locate in majority class, x is a noise sample.

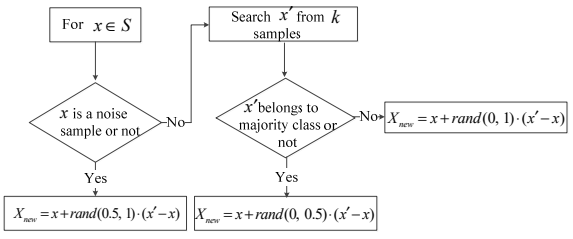

Figure 2: The oversampling steps based on I-SMOTE.

- If x is a noise sample, the oversampling operation may introduce noise into dataset. In order to reduce risk of noise, a sample x' is randomly chosen from minority class samples. Then we generate new samples that close to minority class as follows:

$$X_{new} = x + rand(0.5,\ 1) \cdot (x' - x) \tag{5}$$

- If x is not a noise sample, we search for one sample x' from k samples closest to x, if x' belongs to the majority class, we generate new samples that close to x as follows:

$$X_{new} = x + rand(0,\ 0.5) \cdot (x' - x) \tag{6}$$

if x' belongs to the minority class, we generate new samples as follows:

$$X_{new} = x + rand(0,\ 1) \cdot (x' - x) \tag{7}$$

- Output the new minority samples S'.

4.2 Feature Sorting based on Correlation Analysis

After oversampling, the new dataset can be expressed as minority samples $M' = M \bigcup S'$. The correlation between each feature and classification label is calculated by correlation analysis methods. Detailed steps are as follows:

- For a certain feature vector $X \in \{x_1', x_2', \cdots, x_n'\}^T$ and its corresponding label $Y \in \{0, 1\}$, n is the number of samples in M', their correlation metrics are generated by equations (1)-(3). The correlation metrics can be further standardized and expressed as

$$SU(X, Y) = \frac{I(X;Y)}{H(X) + H(Y)} \tag{8}$$

Thus, the values of correlation metrics between features and labels locate in $[0, 1]$. The value 1 indicates that the feature and label are completely related, and 0 means that they are independent of each other.

- After calculating the correlation, feature ranking is performed. Higher the correlation, more information content it has. It will determine which features in given feature vectors are useful for

classification label. The features with the greatest correlation are at the forefront, and the features with the least correlation are at the last. Thus, the greatest correlation or strongly useful features have high ranking.

4.3 Feature Reduction

In this section, we use the random forests to reduce the sorted features. The goal of feature reduction is to choose as few as features and obtain higher intrusion detection accuracy. Here, we regard the accuracy as a feature fitness. The random forest algorithm is abbreviated as RF. The detailed feature reduction process are as follows. Suppose p is the dimension of the sample. The details processes are shown in Algorithm 1.

Algorithm 1: Feature reduction.

Input: $S0$

Output: $S(t)$ and its dimension $p-t$

For t feature reductions do

Remove the last feature in $S0$ to form $S1$

Compute ACC_{S1}

… …

Remove the last feature to form $S(p-1)$

Compute $ACC_{S0}, ACC_{S1}, ACC_{S2}, \quad ACC_{S(p-1)}$

Out ACC_{St} which is largest

$S(t)$ is chosen with optimal feature

End

4.4 Intrusion Detection Model based on Random Forest

After performing feature reduction on M', the new dataset $S(t)$ is used to train RF classifier to establish the intrusion detection model. Framework of the proposed intrusion detection system consists of three steps: Oversampling, feature reduction and test model.

Step1: Oversampling.

The purpose of oversampling is to oversample minority samples, and reduce the imbalance between minority samples and majority samples.

Step2: Feature Reduction.

The feature reduction algorithm is applied to the oversampled dataset to generate the new transformed dataset for testing.

Step3: Test Model.

Apply the new transformed dataset from step1 and step 2 to train RF classifier and obtain the intrusion detection model. New testing samples are brought into the intrusion detection model to test the performance of the model.

5 EXPERIMENTAL SETTING

5.1 Dataset Description

The dataset used in the paper is based on the NSL-KDD dataset, which is a modified version of the KDD Cup 99 (Tavalaee et al., 2009) dataset. The KDD Cup 99 dataset contains 494,020 samples. Each sample represents a TCP connection records represented by a 41-dimensional feature vector, in which 9 nominal features and 32 continuous features. Each category of the dataset is labeled as one out of five classes, which are normal traffic and four different classes of attacks, i.e., Probing, DOS, R2L, U2R. Also, the NSL-KDD is an imbalance dataset in which the number of DOS is about 10000 times than U2R.

Although the KDD Cup 99 dataset is widely applied in intrusion detection field, the dataset has some flaws, for example, there exist many duplicated samples, leading to classifiers trained with it to be biased toward the samples with larger number. Thus, (Tavalaee et al., 2009) proposed a more complete dataset, namely, the NSL-KDD dataset. The optimization makes the NSL-KDD dataset more reasonable in data distribution. (Bhattacharya et al., 2015), D (De La Hoz et al., 2015) and (Kim et al., 2014) illustrated that the NSL-KDD dataset can be regarded as an ideal dataset in intrusion detection field.

5.2 Experimental Setup

The empirical experiments in our work were all implemented on a computer with an Intel Core i7-7700

CPU @ 3.60GHz with 16.0 GB RAM running Windows 10. The feature reduction, transformation and RF classifier test were run using Python.

The dataset is divided into two disjoint parts, which are respectively used as the training and the testing. The 10-fold cross validation method was applied to train and test the proposed classifier by using training dataset. In this method, the dataset is divided into 10 un-duplicated subsets, and any nine of ten are used for training and the remaining one for testing. Thus, after running 10 times, each subset of the initial dataset has an equal opportunity to be selected as a training or testing. Thus, the RF classifier will be trained and tested 10 times. Finally, the performance of intrusion detection is evaluated by the testing dataset.

5.3 Experimental Results

This section shows the benefits of the proposed intrusion detection model for multi-class classification problem. In this work, we consider the rates of detection, false alarms and accuracy, where the rates of detection, false alarms and accuracy are widely applied in related work to indicate the performance of intrusion detection model. To verify the effectiveness of this model, we show the experiment from four aspects. First, we explain why we choose the random forest as the multi-class classifier. Second, the oversampling operation improves performance of the classifier is confirmed. Third, we show advantages of our scheme in terms of training and testing time compared with other classic schemes. Finally, to further investigate the advantages of the proposed model, a comparison of overall efficiency is illustrated between our proposed model and other related methods.

The experimental results shown in Table 1 and Table 2 illustrate advantages of random forest compared with SVM, GBDT, LRL1 and LRL2, where LRL1 and LRL2 represent the logistic regression with L1 regularization and L2 regularization respectively.

Table 1: The detection rate obtained by different classifier in 5 classification detection.

Classifier Types	Probe	DOS	U2R	R2L	Normal
SVM	0.952471	0.999629	0.230769	0.183919	0.993079
RF	**0.996139**	**0.999695**	**0.519231**	**0.958793**	**0.999584**
GBDT	0.940803	0.992862	0.269231	0.688442	0.997015
LRL1	0.932481	0.998492	0.346153	0.358794	0.984778
LRL2	0.929993	0.998524	0.230769	0.098492	0.983011

Table 2: The overall performance of different classifier in 5 classification detection.

Classifier Types	ACC	Detection Rate	False Alarm Rate
SVM	0.985003	0.975728	0.006921
RF	**0.993785**	**0.997867**	**0.000416**
GBDT	0.987561	0.976701	0.002985
LRL1	0.979731	0.973937	0.015222
LRL2	0.976464	0.968941	0.016989

As shown in Table 1, we can see that the random forest has the highest detection rate, especially in detecting U2R and R2L attacks. For U2R, the detection rates of SVM, GBDT, LRL1 and LRL2 are 0.23, 0.27, 0.35 and 0.23 respectively. Then, the detection rate of RF is 0.52 which significantly outperforms other classifiers. Also, for R2L, the detection rates of SVM, GBDT, LRL1 and LRL2 are 0.18, 0.69, 0.36 and 0.098 respectively. Then, the detection rate of RF is 0.96.

The overall performance of different classifier is illustrated in Table 2. RF can produce the highest ACC and detection rate, and lowest false alarm rate. The ACC of SVM, GBDT, LRL1 and LRL2 are 0.985, 0.987, 0.979 and 0.976, and RF is 0.994 which is larger than other values. The overall detection values of SVM, GBDT, LRL1 and LRL2 are 0.975, 0.976, 0.974 and 0.969, and RF is 0.997 which is also larger than other values. The most obvious advantage of random forest is reflected in false alarm rate. The false alarm rate of SVM, GBDT, LRL1 and LRL2 are 0.0069, 0.0029, 0.0152 and 0.0169, and RF is 0.0004 which is nearly 10 times smaller than other values.

The experimental results shown in Table 3 and Table 4 illustrate advantages of the preRF compared with preSVM, preGBDT, preLRL1 and preLRL2, where preRF indicates the combination of oversampling, feature reduction and random forest, preSVM indicates the combination of oversampling, feature reduction and SVM, preGBDT indicates the combination of oversampling, feature reduction and GBDT, preLRL1 indicates the combination of oversampling, feature reduction and LRL1, and preLRL2 indicates the combination of oversampling, feature reduction and LRL2.

As shown in Table 3, we can see that our method has the highest detection rate in five classification, especially in detecting U2R attacks. For U2R, the detection rates of preSVM, preGBDT, preLRL1 and preLRL2 are 0.46, 0.60, 0.59 and 0.56 respectively. Then, the detection rate of preRF is 0.96 which

Table 3: The detection rate obtained by different combination method.

Classifier Types	Probe	DOS	U2R	R2L	Normal
preSVM	0.955387	0.999673	0.458042	0.934307	0.996864
preRF	**0.996654**	**0.999738**	**0.963286**	**0.997168**	**0.999683**
preG-BDT	0.949011	0.994316	0.596153	0.929009	0.997065
preLRL1	0.933537	0.998543	0.590909	0.931841	0.985142
preLRL2	0.930934	0.998671	0.562937	0.936683	0.989411

Table 4: The overall performance of different combination method.

Methods	ACC	Detection Rate	False Alarm Rate
preSVM	0.986988	0.977365	0.003136
preRF(Our method)	**0.999088**	**0.998508**	**0.000317**
preGBDT	0.988894	0.973033	0.002935
preLRL1	0.979315	0.973637	0.014858
preLRL2	0.981514	0.973818	0.010589

significantly outperforms other classifiers. The overall performance of different combination classifier is illustrated in Table 4. The ACC of preSVM, preGBDT, preLRL1 and preLRL2 are 0.987, 0.989, 0.979 and 0.981, and preRF is 0.999 which is larger than other values. The overall detection rate of preSVM, preGBDT, preLRL1 and preLRL2 are 0.977, 0.973, 0.974 and 0.974, and preRF is 0.998 which is also larger than other values. The most obvious advantage of random forest is reflected in false alarm rate. The false alarm rate of preSVM, preGBDT, preLRL1 and preLRL2 are 0.0031, 0.0029, 0.0148 and 0.0106, and preRF is 0.0003 which is nearly 10 times smaller than preSVM and preGBDT, 100 times smaller than preLRL1 and preLRL2. In summary, the preRF is good at detection the Probe, DOS, U2R, R2L and Normal classes.

The advantages of oversampling and feature reduction can be illustrated in Table 2 and Table 4. In Table 4, preRF provides an ACC, detection rate, and false alarm rate of 0.999088, 0.998508, and 0.000317 respectively, results which are better or similar to the RF classifier shown in Table 2, which for 0.993785, 0.997867 and 0.000416 respectively. The other methods in Table 4 and Table 2 also achieve the same effect.

The advantages of our scheme in terms of training and testing time compared with other classic schemes are shown in Table 5. The training time of preSVM is about 306s, preRF is about 10s, preGBDT is about 38s, preLRL1 is about 59s, preLRL2 is about 20s. The method proposed in this paper has minimal training time. At the same time, the testing time of preSVM is about 1ms, preRF is about 11ms, preGBDT is about 1ms, preLRL1 is about 1ms, preLRL2 is about 1ms. Although preRF has the longest testing time, it has the shortest sum time from the whole processes.

Table 5: The training time and testing time of different combination method.

Classifier Types	Training time	Testing time	Sum time
preSVM	306s	1ms	306.02s
preRF	**10s**	**11ms**	**10.18s**
preGBDT	38s	1ms	3 8.02s
preLRL1	59s	1ms	59.02s
preLRL2	20s	1ms	20.02s

Regarding the previous results, we further examine the proposed method over the five classes. Performance of different methods is illustrated in Fig.3. The CANN (6), CANN (19), k-NN (6) and k-NN (19) are described in detail in (Wei-Chao et al., 2015), where 6 and 19 represent the dimensions of CANN and k-NN with feature reduction, and preRF (30) indicates that the optimal dimension is 30 in our method after feature reduction .To detect the normal, probe and Dos classes, the performance of preRF is not obvious compared with CANN and k-NN. However, preRF can correctly detect U2R and R2L with the highest detection rate. These results indicate that the classification with oversampling and feature reduction is suitable for detection of U2R and R2L in the five-classification problem.

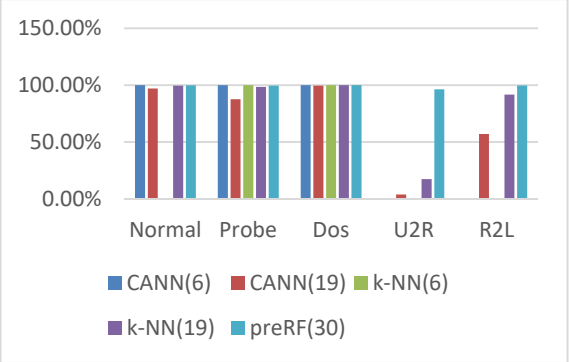

Figure 3: The performance of preRF, CANN and k-NN.

6 CONCLUSIONS

This paper proposes a multi-classification method that applies the improved synthetic minority oversampling technique (I-SMOTE) to balance the dataset, employs correlation analysis and random forest to reduce features and uses the random forest algorithm to train the classifier for multi-attack type detection. The experimental results based on the NSL-KDD dataset show that it achieves a better and more robust performance in terms of accuracy, detection rate, false alarms and training speed.

ACKNOWLEDGEMENTS

The authors would like to thank the editorial board and reviewers. This work was supported by the Research on Key Technologies of High Security and Trustworthy Mobile Terminal Operating System Security Protection (2017YFB0801902).

REFERENCES

Amin Hatef M., Shaker V., Reza Jabbarpour M. et al. HIDCC: A hybrid intrusion detection approach in cloud computing. Concurrency Computat: Pract Exper., 2018; 30: e4171.

Luo B., Xia J., A novel intrusion detection system based on feature generation with visualization strategy. Expert Syst. Appl. 41 (9) (2014) 4139-4147.

Tan L. L., Li C. and Xia J. M., Application of self-organizing feature map neural network based on K-means clustering in network intrusion detection. CMC-Computers Materials & Continua, 61 (1) (2019) 275-288.

Bamakan S. M. H., Wang H., Tian Y., Shi Y., An effective intrusion detection framework based on mclp/svm optimized by time-varying chaos particle swarm optimization, Neurocomputing 199 (2016) 90-102.

Kuang F., Xu W., Zhang S., A novel hybrid kpca and svm with ga model for intrusion detection. Appl. Soft. Comput. 18 (4) (2014) 178-184.

Horng S. J., Su M. Y., Chen Y. H., Kao T. W. „ A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert Syst. Appl. 38 (1) (2011) 306-313.

Li X. K., Chen W., and Zhang Q. R., 2020. Building Auto-Encoder Intrusion Detection System based on random forest feature selection. 95.10.1016/j.cose.2020.10 1851.

Chauhan,V. Kumar, S. Pundir, and E. S. Pilli. 2013. A comparative study of classification techniques for intrusion detection. In 2013 International Symposium on Computational and Business Intelligence.40–43.

Amira Sayed A. Aziz, Sanaa EL-Ola Hanafi, and Aboul Ella Hassanien. 2017. Comparison of classification techniques applied for network intrusion detection and classification. Journal of Applied Logic 24 (2017), 109–118.

Manjula C. Belavagi and Balachandra Muniyal. 2016. Performance evaluation of supervised machine learning algorithms for intrusion detection. Procedia Computer Science 89(Jan.2016),117–123

Golrang A., Golrang A. M., and Yayilgan S. Y.. 2020. A Novel Hybrid IDS Based on Modified NSGAII-ANN and Random Forest. 9 (4) 2020.

Tesfahun A. and Bhaskari D. L., 2013. Intrusion detection using random forests classifier with SMOTE and feature reduction In 2013 International Conference on Cloud Ubiquitous Computing Emerging Technologies.127–132.

Malik A. J., Shahzad W., and Khan F. A., 2011. Binary PSO and random forests algorithm for PROBE attacks detection in a network. In 2011 IEEE Congress of Evolutionary Computation (CEC'11).662–668.

Malik A. J. and Khan F. A., 2013. A hybrid technique using multi-objective particle swarm optimization and random forests for PROBE attacks detection in a network. In 2013 IEEE International Conference on Systems, Man, and Cybernetics. 2473–2478.

Arif Jamal Malik, Waseem Shahzad, and Farrukh Aslam Khan. 2015. Network intrusion detection using hybrid binary PSO and random forests algorithm.8, 16 (Nov. 2015), 2646–2660.

Zhong SH, Huang HJ, Chen AB. 2011. An effective intrusion detection model based on random forest and neural networks. Advanced Materials Research, vol. 267 (308), pp. 308–313.

Md. Al Mehedi Hasan, Mohammed Nasser, Biprodip and Shamim Ahmad. 2014. Support Vector Machine and Random Forest Modeling for IDS, JILSA, pp. 45–52.

Tan X. P., Su S. J., Huang Z. P. , Guo X. J., Wireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm, Sensors 2019. 19, 203.

Wang L., Chen H. M., Unbalanced dataset classification method based on NKSMOTE algorithm. Computer Science, 2018, 9 (45), 260-265.

Tavalaee M., Bagheri E., Lu W., Ghorbani A. A., 2009. A detailed analysis of the kdd cup 99 data set, in: Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

Bhattacharya S., Selvakumar S., 2015. LAWRA: a layered wrapper feature selection approach for network attack detection, Security Commun Netw 8 (18), 3459–3468.

De La Hoz E., Ortiz A., Ortega J., Prieto B., 2015. PCA filtering and probabilistic som for network intrusion detection, Neurocomputing 164, 71–81.

Kim G., Lee S., Kim S., 2014. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, Expert Syst. Appl. 41 (4), 1690–1700.

Wei-Chao L., Shih-Wen Ke, Chih-Fong T., CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-Based Systems. 78 (2015) 13-21.

Chen Tongbao, Wen Liangming, Li Jianhui. A Data Prediction Method Based on Feature Selection and Transfer Learning[J].Frontiers of Data & Computing,2020,2(2): 145-154.