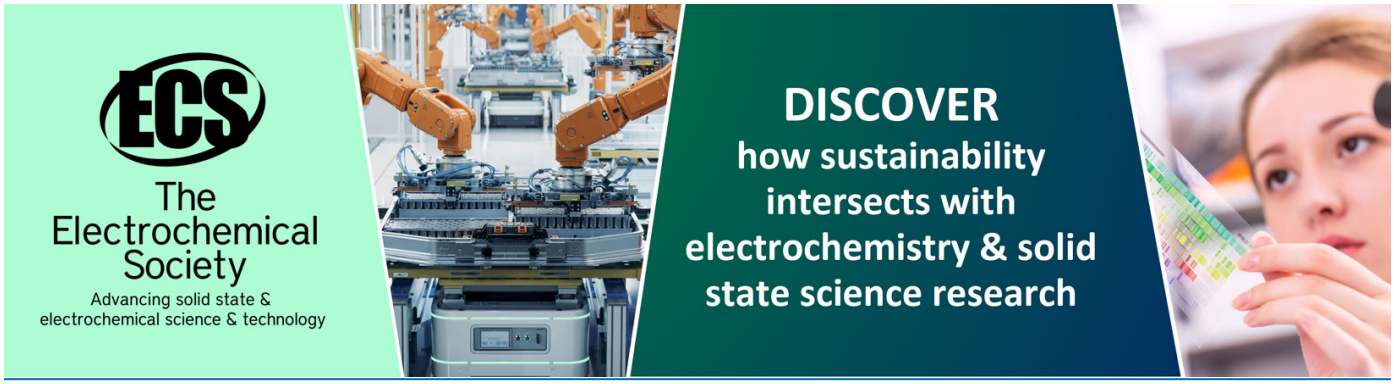
Research on intrusion detection method based on feature selection and integrated learning



Research on intrusion detection method based on feature selection and integrated learning

Qun Liu[1,a], Zhiyong Tong[2,b], Shuiqing Wang[1,c], Ziheng Yang[1,d*]

[1]Electronic Engineering College, Heilongjiang University, Haerbin, Heilongjiang, China

[2]Heilongjiang Provincial Military Command, Haerbin, Heilonhjiang, China

[a]email: lqun9601@163.com, [b]email: tzy1997@163.com,

[c]email: 17863934828@163.com,

*Correspondence: [d]email: yzh@hlju.edu.cn

Abstract. With the introduction of computer technology, network attacks have become more frequent. Some illegal elements may intrude into computers through network attacks to tamper with messages, spread viruses and other destructive behaviors, causing great damage to personal sensitive information, industrial control networks, transaction systems, etc. . For this, this design proposes an improved intrusion detection method based on feature selection and integrated model. The NSL-KDD training data set is used to evaluate the proposed model. First, balance the data categories through the SMOTE-ENN method, and then use feature selection technology and PCA feature extraction technology to reduce the number of irrelevant features and improve the classification accuracy. Finally, using CART as the base classifier, Bagging technology is used to establish an integrated model and an intrusion detection system. Experimental results show that the CART-based Bagging method provides better accuracy, lower false alarm rate and faster model training speed, and the system can detect intrusion attacks with similar attributes and has a certain degree of adaptability.

1. Introduction

With the rapid development of information and network technology, industrial equipment has formed a network, and the high value or absolute number of various types of transactions have caused industry and commerce and their customers to face serious risks. These risks are exacerbated by the increase in the number and maturity of interconnected networks and malicious opponents. Ensuring information security is an important task that must be completed in today's rapid development of the Internet of Things.

Intrusion Detection System (IDS) is an important tool for computer network protection. The potential of intrusion detection technology has a fundamental impact on the performance of intrusion detection systems. IDS prevents possible intrusions by detecting network traffic to ensure the security and integrity of the network information[1]. Recently, IDS systems based on machine learning (ML) have been widely used in network intrusion detection tasks[2]. This article turns to machine learning algorithms for intrusion detection. Machine learning can allow the system to automatically improve performance as experience increases. The decision tree classification algorithm uses the probability of various events to construct a complete decision tree, which is used to assess and predict event risks. The decision tree classification algorithm satisfies the needs of intrusion detection systems very well.

It can not only classify attacks and abnormal behaviors, but also actively detect and prevent risks. For the task of intrusion detection, this paper proposes an improved intrusion detection method based on feature selection and integrated model. This method can improve the recognition rate of network attacks and reduce the false alarm rate to a certain extent.

2. Intrusion detection system design

This paper is oriented to the task of intrusion detection. On the NSL-KDD data set, the CART decision tree is used as the base classifier to construct an intrusion detection mechanism through the integrated learning method in machine learning.

2.1. Decision tree

Decision tree is a classifier. The main advantage of decision tree over many other classification techniques is that they generate a set of simple and efficient rules that can be combined with other network protection methods. The core key point affecting the learning of decision tree is how to find the optimal division attribute, we hope that as the splitting process progresses, the data contained in the nodes in the decision tree belongs to the same category to the greatest extent, and then introduces the CART decision tree[3] classification algorithm used in this article.

CART algorithm: CART algorithm is also a decision tree algorithm, and the division of its nodes is based on the Gini index formula (1)

$$Gini(D)=\sum_{k=1}^{|y|}\sum_{k'\neq k}p_k p_{k'}=1-\sum_{k=1}^{|y|}p_k{}^2 \tag{1}$$

where *Gini(D)* means extracting two samples from data set *D*. The Gini exponent for attribute a can be defined as the formula (2)

$$Gini\_index(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v) \tag{2}$$

Finally, in the attribute set A, the attribute with the smallest Gini index after the division is used as the optimal division basis as shown in formula (3)

$$a_*=\arg\min Gini\_index(D,a) \tag{3}$$

where $a\in A$.

2.2. Intrusion detection based on integrated learning

2.2.1. Data balance processing.

There have been many researches on intrusion detection based on machine learning methods, and the most common data used among them is the NSL-KDD dataset[4]. This article uses KDDTrain+ and KDDTest+. The data distribution is shown in Table 1.

It can be konwn from Table 1 that although the NSL-KDD data set has been processed, its category distribution is obviously unbalanced, of which Normal category data accounts for 53.46%, while U2R only has 0.04%. In a real network environment, the normal category of network traffic is also much higher than the attack category. In this case, IDS will focus on the monitoring of normal data flow, thereby improving the accuracy of the overall data flow recognition. Therefore, some attack data traffic may be ignored, which reduces the performance of the protection system. At present, sampling methods such as over-sampling and under-sampling are commonly used to deal with the problem of unbalanced data. Over-sampling methods repeat or synthesize new samples for insufficient classes, while under-sampling will reduce samples of larger classes. In this article, we first use the over-sampling method SMOTE to amplify samples with fewer sample categories[5], and finally use the under-sampling method ENN to remove noise samples, thus forming the SMTE-ENN method.

Table 1 Distribution of cyber attacks

Data Set	Number Statistics					
	Total	Normal	Dos	Probe	U2R	R2L
Training Set	125973	67343	45927	11656	52	995
		53.46%	36.46%	9.25%	0.04%	0.79%
Test Set	22544	9711	7458	2421	200	2654
		43.08%	33.08%	10.74%	0.89%	12.22%

As shown in Figure 1, The sample sizes of Normal, Dos, Probe, R2L, and U2R categories in the original data are 67343, 45927, 11656, 995, and 52, respectively. After the SMOTE-ENN method, the sample sizes are 66877, 67285, 67269, 67277 and 67333, we will conduct research on intrusion detection methods based on decision tree on the basis of this sample.

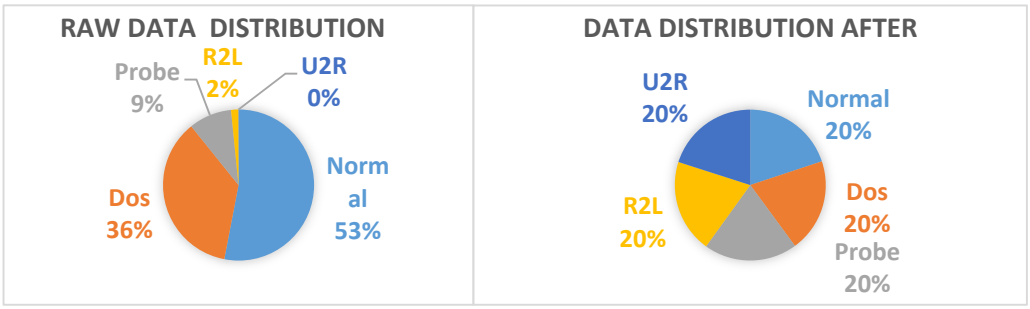


Fig 1 The distribution of training data

2.2.2. PCA feature extraction.

Feature extraction has a wide range of applications in data mining tasks. Generally, before this step is applied to training the model, in order to make the model easier to train, we will adjust the dimensionality of the feature space, generally using dimensionality reduction methods. Feature extraction can not only reduce the dimension of the feature space and the complexity of the task, but also reduce the redundancy of data information, thereby improving the performance of the model. In the face of intrusion detection tasks, this paper introduces the principal component analysis (PCA) algorithm[6] to extract data features.

The PCA algorithm implementation process is as follows

(1) Let input data $X=\{x^1,x^2,....,x^n\}$,where X is n p-dimensional column vectors.

(2) For data centralization, such as formula (4) and formula (5)

$$\mu=\frac{1}{N}\sum_{t=1}^{N}x^t \tag{4}$$

$$x^i=x^i-\mu \tag{5}$$

where $i\in\{1,2,...,n\}$.

(3) Calculate the covariance matrix formula (6)

$$C=\frac{1}{N}XX^T \tag{6}$$

(4) Decompose the eigenvalue formula of covariance matrix C (7)

$$C_{v_i}=\lambda_i v_i \tag{7}$$

where $\lambda_i(i=1,2,...,p)$ is the eigenvalue, and its corresponding eigenvector is $v_i(i=1,2,...,p)$. The eigenvectors corresponding to the first k eigenvalues are used to form the projection matrix $A=\{v_1,v_2,...,v_k\}$, where v_i is the eigenvector corresponding to λ_i .

(5) Finally, project the original sample to the new feature space to obtain the new dimensionality reduction sample formula (8)

$$X' = W^T X \tag{8}$$

Where W^T is a trainable matrix, and X' is *n p*-dimensional column vectors.

2.2.3. Data balance processing.

At present, the integrated learning method is often used in data min-ing tasks. Ensemble learning is a method that uses multiple classifiers. Its purpose is to improve the robustness of the model and learn how to improve the classification performance from each subclass-ifier. This method subdivides a big problem, and then adopts targeted solutions for each part. The suc-cess of the ensemble method depends on the diversity of the misclassified instances of each subclass-ifier[7]. This paper is oriented to intrusion detection tasks and uses CART as the classifier to constr-uct a Bagging ensemble learning model.

First introduce the Bagging method of building an integrated CART decision tree. Under this method, each decision tree is independently trained through the bootstrap method, and then they will be aggregated through an appropriate method. Usually, we divide the training set into T subsets, and then use each subset to build a decision tree, and finally these independent decision tree are built into an integrated decision tree. From the statistical facts, we need to use diversified training samples to obtain better integrated learning results and improve model performance. For this reason, we often use Bootstrap technology, which constructs T subsets through repeated sampling with replacement. The sample x in each subset may appear repeatedly or not. Finally, each subset will be used to train a specific decision tree.

After the training is over, we use the voting mechanism to combine the predicted output, and the process is shown in Algorithm 1.

Algorithm1 Bagging Algorithm Flow

Input: Training set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$;

Base learner $\mathcal{L}$;

Number of training rounds T.

Output: $H(x) = \arg\max \sum_{t-1}^{T} \prod (h_t(x) = y)$

Begin:

1. **for** *t*=1,2,…,T **do**
2. $h_t = \mathcal{L}(D, D_{bs})$
3. **end for**

3. Results

This article divides the data set into two types of normal and abnormal data according to sample labels. The following first introduces the classification indicators used: false negative (FN), false positive (FP), true positive (TP) and true negative (TN).

This paper adopts the *accuracy* rate as the formula (9) and the false alarm rate *FAR* as the formula (10) as the evaluation indicators.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$FAR = \frac{FP}{FP + TN} \tag{10}$$

This paper sorts the importance of features by Gini coefficients, as shown in Figure 2, and the first 35 features are selected to be studied separately, aiming to find a subset that can work with a classifier to produce higher classification accuracy. We use the CART decision tree as the base classifier and use the BAGG integration technology to build the integration model. Using a subset of the NSL-KDD dataset, the 35 feature dataset trained and tested these models. The training and testing phases are performed on two separate datasets (KDDTrain+ and KDDTest+). This paper uses Scikit-learn as the basic tool for algorithm implementation.

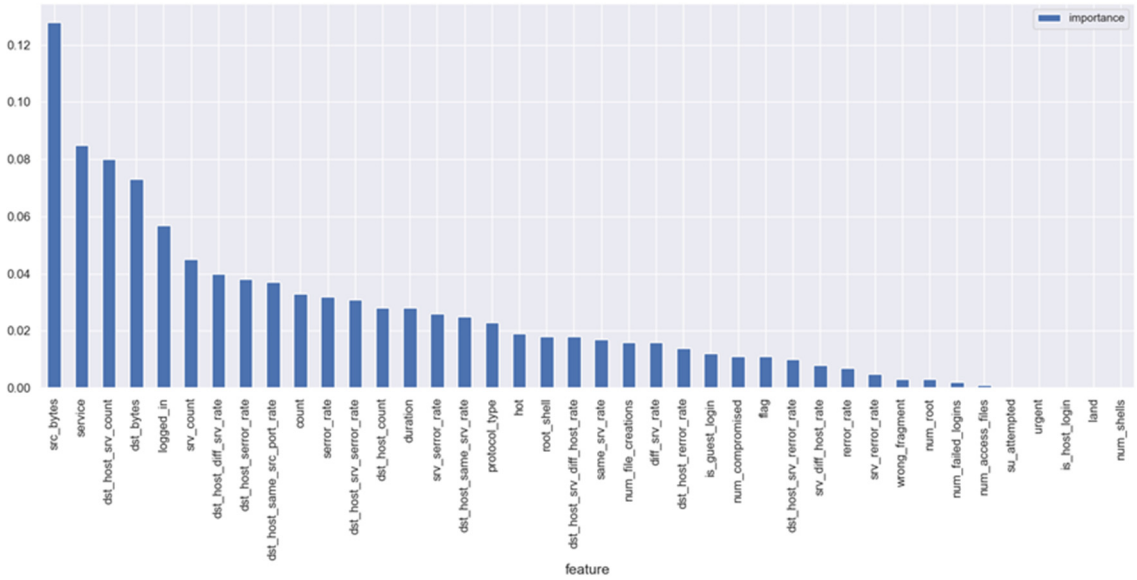


Fig 2 Sort feature importance by Gini

Table 2 shows the experimental results of the ensemble model and other methods proposed in this paper using 35 feature subsets to train and test.

Table 2 experimental results with 35 features

Model	Accuracy(%)	FAR(%)
SVM	80.03	3.88
KNN	78.94	3.01
CART	80.23	3.19
Naïve Bayes	80.87	8.12
Bagging(CART DT)	84.68	1.81

In total, by selecting 35 feature subsets, the classification performance of the Bagging (CART DT) model was greatly improved, and its accuracy increased from 80.51% to 84.68%, and its false positive rate decreased from 2.09% to 1.81%. This is because the Gini coefficients used in the selection feature can also be used as a partition attribute in the decision tree. Therefore, using Gini coefficients for feature selection can help improve the performance of the Bagging (CART DT) model. This makes the Bagging (CART DT) model perform better than other methods on a subset of 35 features. It is worth noting that although some methods currently have a higher accuracy rate than the Bedding (CART DT) model we designed, their false positive rate is also higher than the model we designed. Therefore, the intrusion detection model in this article should not only improve the recognition rate of network

attacks, but also reduce the false alarm rate as much as possible to improve the overall performance of IDS.

4. Conclusion

In general, this paper proposes an improved intrusion detection method based on feature selection and integrated model. The NSL-KDD training data set is used to evaluate the proposed model. First, balance the data categories through the SMOTE-ENN method, and then use feature selection technology and PCA feature extraction technology to reduce the number of irrelevant features and improve the classification accuracy. Using CART as the base classifier, Bagging technology is used to build an ensemble model. The experimental results show that the method we designed has a higher recognition rate and a lower false positive rate for network attack data.

References

[1] Hoque M S, Mukit M, Bikas M, et al. An implementation of intrusion detection system using genetic algorithm[J]. arXiv preprint arXiv:1204.1336, 2012.
[2] Biswas S K. Intrusion detection using machine learning: A comparison study[J]. International Journal of pure and applied mathematics, 2018, 118(19): 101-114.
[3] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics, 1991, 21(3): 660-674.
[4] Tavallaee M, Bagheri E, Lu W, et al. A detailed analysis of the KDD CUP 99 data set[C]. 2009 IEEE symposium on computational intelligence for security and defense applications. IEEE, 2009: 1-6.
[5] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.
[6] Hadri A , Chougdali K , Touahni R . Intrusion detection system using PCA and Fuzzy PCA techniques[C]. 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS). IEEE, 2016.
[7] Aravind M , Kalaiselvi V . Design of an intrusion detection system based on distance feature using ensemble classifier[C]. International Conference on Signal Processing. 2017:1-6.