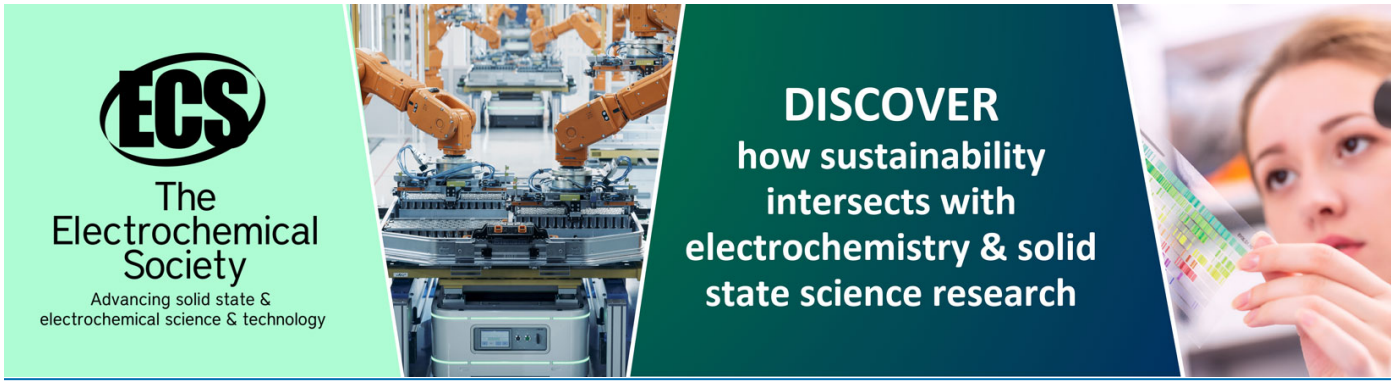
Researching on Multiple Machine Learning for Anomaly Detection



Researching on Multiple Machine Learning for Anomaly Detection

Yuanyuan Sun [1, 2, 3a], Yongming Wang [1, 2b], Lili Guo [3c], Zhongsong Ma[3], Shan Jin[3] and Huiping Wang[3]

[1]School of Cyber Security, University of Chinese Academy of Sciences, Beijing China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing China
[3]Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Key Laboratory of Space Utilization, Beijing China

Corresponding author's e-mail address: [a]sunyuanyuan@csu.ac.cn, [b]wangyongming@iie.ac.cn, [c]guolili@csu.ac.cn

Abstract. Firstly, we introduce intrusion detection system and anomaly detection. And then we do some research on machine learning techniques for anomaly detection by network dataset NSL-KDD. The machine learning algorithms such as J48, Random forest, SVM, Vote, Stacking are selected. Random Forest, Vote and stacking are ensemble learning methods. We try to test and verify performance of multiple machine learning methods on a 20 per cent NSL-KDD dataset by experiment. The experiment data has two parts. First, the 20 per cent NSL-KDD dataset is classified into normal and anomaly. Second, the feature of attack type is added to the 20 per cent NSL-KDD dataset, and then a new dataset is generated. It is classified into normal and other four classes of attack. The experiment is accomplished by WEKA. The result is compared on the basis of typical indexes and confusion matrix. At last,we can draw a conclusion that an appropriate ensemble classifier can achieve better classification performance than a single classifier for anomaly detection .

1. Introduction

Recently, with the mushroom growth of computer science and communication network, the issue of security is becoming more and more important not only in Computer Network, but also involve of Mobile Telecommunication Networks, Wireless Sensor Networks, and some other industrial process etc. In this paper, we discuss a kind of security technique that is anomaly detection. It is a part of IDS-Intrusion Detection System.

The intrusion detection system can monitor network transmission and issue an alert for suspicious activities. When it detects the threat, it can give report of suspicious activity to the network server or the network administrator. Under some circumstance, the IDS may also take positive action to prevent the intrusion, such as blocking source IP address or the users which accessing the network according to anomalous or malicious traffic.

The technology of intrusion detection system can be divided into the Misuse detection and Anomaly detection [1].

Misuse detection IDS can also be seen as signature based detection. Firstly Misuse detection attempts to model abnormal behaviour, subsequently it will monitor activities of the program on the network and extract features from those activities, comparing the features with the malware features in

a database that come from known malicious threats. This is something like a way that virus detection software detects viruses. The defect is that there will be delaying time between the new features of suspicious behaviour and the features for detecting the menace which has been applied to the IDS early. The IDS has no ability to detect the new menace during the delaying time.

The anomaly based IDS attempts to model normal behaviour, which will set up normal matrix as the baseline. The elements of matrices derived from various indicators of the system. Such as CPU utilization, memory utilization, time and number of logins, network activity, file exchanges, bandwidth, protocols, etc. If the behaviour is significantly different the baseline, anomaly detection IDS will alert the system to defend.

In our paper, we will discuss the method of anomaly detection of IDS. We plan to find high performance machine learning techniques for identifying anomaly information from a public network intrusion dataset.

2. Related work

S.Revathi et al. applied several single machine learning techniques such as J48, Random Forest, SVM, CART, Navie Bayes on NSL-KDD Dataset for intrusion detection[2]. They found a best one that is Random Forest in their experiment.

Homoliak used well known NSL-KDD 1999 dataset to demonstrate various convergence optimization experiments of a back- propagation artificial neural network [3].

Aalahi proposed a hybrid method that is consisted of SVM and genetic algorithm which is used for intrusion detection [4]. The hybrid algorithm proposed can decrease the quantity of features from 45 to 10. The features are classified into three grade adopting GA algorithm. The results show that the proposed hybrid algorithm achieves 0.973 as the true-positive value and 0.017 as the false-positive value.

Alotaibi utilized a few machine learning algorithms and seek out some hopeful ones to enhance the accuracy on the public dataset [5]. That is Bagging, Random Forests and Extra Trees. Then they adopted a majority voting method to vote the prediction of those algorithms. They found that both customized voting method and bagging classifier method get good outcome.

3. Classifier

In this paper, the classifiers we selected include J48, Random forest, SVM(Support Vector Machine),Vote, Stacking.

3.1. J48

J48 algorithm derives from ID3 and c4.5 algorithm which is been modified. It is implemented by WEKA (Waikato for Knowledge Analysis) software and widely used in building decision trees [6].

In decision tree, the internal nodes represent test on attribute, branch denotes result of test, and leaf nodes signify class labels. The route from leaf to root is called classifying rule. The tree is consisted of chance node, decision node and end nodes.

The standard tree is consisted of a root, several of branches, leaves and end nodes. There is an estimation criterion for every node in decision tree which is used to select relevant input variable for predicting. The estimation criteria, which is based on information gain and entropy reduction, is used to identify input variables.

The E (entropy)is defined by following formula:

$$E=-pP\times\log2(pP)-pN\times\log2(pN) \quad (1)$$

pP -- proportion of positive examples
pN -- proportion of positive examples

3.2. Random forest

Breiman introduced RF - the Random Forests classifier in 2001 [7]. It is widely focused by many researchers in recent years. RF is a kind of statistical learning theory which utilizes re-sampling

method to extract multiple samples from the original one. The bootstrap sample of training set is adopted to train every tree in the forest. And then use the method of voting to decide the output of model trees. RF is extension of bagging classification tree. It is an ensemble of trees. Both empirical and theoretical have shown that RF has very high precision, which has a good tolerance to both noise and outliers. And it is not easy to over fitting. In a manner of speaking, RF is a natural nonlinear modelling tool.

3.3. Support Vector Machine (SVM)

SVM is an available statistical theory and this technology can be used to solve engineering problems as well. SVM has outstanding generalization performance which was firstly proposed by Vapnik [8]. It is a popular method to solve classification problem in fact. In the statistical learning theory, SVM comes from Structural Risk Minimization principle (SRM). The Structural Risk Minimization means maximizing the distance (margin) between different classes. A SVM algorithm will construct a hyper-plane. This hyper-plane has the largest distance to any class in multi-dimensional space. It is illustrated in Figure 1.

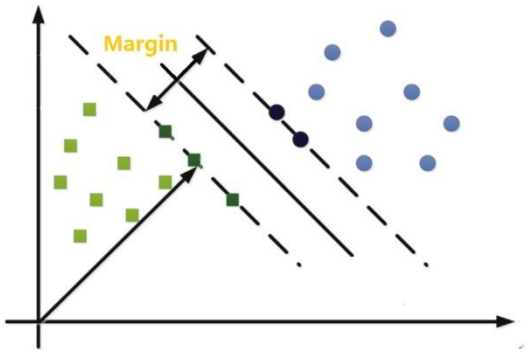


Figure 1. SVM classifier using hyperplane

3.4. Vote

Vote is a kind of ensemble learning method. Sometimes we can get some advantages by using vote to combine several classifiers. For example, high performance of generalization, better accuracy and increasing robustness can be achieved. The process of vote can be illustrated in Figure 2. There are several combinational voting methods such as Majority voting, Average of probabilities, Minimum probabilities, Product of probabilities, Maximum probabilities. Majority voting which can be said that is one of the most popular technologies.

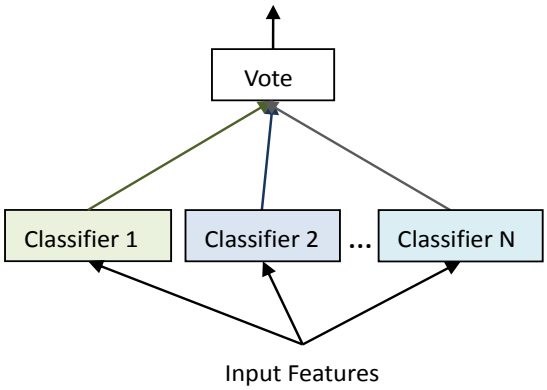


Figure 2. Architecture of Vote

3.5. Stacking

The stacking method is illustrated in Figure 3. Sometimes stacking can be seen as stacked generalization,which involve of training the learning algorithms by combining the results of some

other learning algorithms. In the first place, all the base models are trained by the available data, and then a Meta model is trained to make a terminal decision by all the predictions of the base models.

4. Methodology of Experiment

4.1. System Framework

The experiment is conducted by WEKA -Waikato Environment for Knowledge Analysis. WEKA is freely available software. WEKA provides implementation of machine learning algorithms. The developer can perform clustering, classification, visualization and association by WEKA [9].

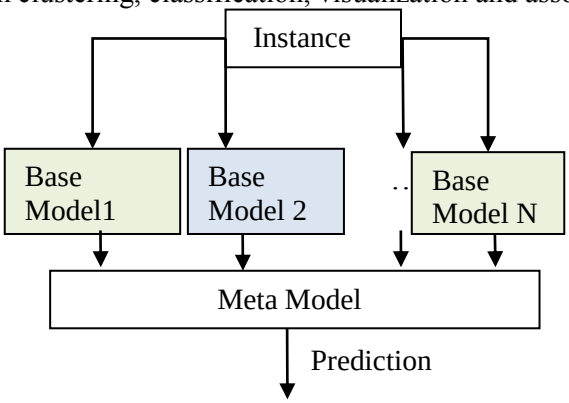


Figure 3. Architecture of Stacking

In this paper, we choose high performance single classifier and research on different ensemble methods by combining those single classifiers together in order to enhance the performance of anomaly detection. For this purpose, we choose three default setting classifiers in WEKA as base classifiers such as J48, Random forest, SVM. The methods of 'Vote' and 'Stacking' are selected as ensemble methods to accomplish the experiment. 'Vote' and 'Stacking' are under the directory of Meta classifiers in WEKA.

The entire process is depicted in Figure 4. Concretely, the system framework is divided into the following stages:

- Data pre-processing. We choose the filter of 'Normalize' to conduct preprocess.
- Data classification with 'J48'.
- Data classification with 'RF'
- Data classification with 'SVM'
- Data classification with 'Vote_ 1' based on J48, RF, and SVM. The combinatorial rule is Average of probabilities.
- Data classification with' Vote_ 2' based on J48, RF, and SVM. The combinatorial rule is Majority voting.
- Data classification with 'Stacking_1'. The base classifiers are J48, RF, and the meta classifier is logistic regression.
- Data classification with 'Stacking_2'. The base classifiers are J48, RF, and the meta classifier is SVM.
- Data classification with 'Stacking_3'. The base classifiers are J48, RF, and the meta classifier is J48.
- Data classification with 'Stacking_4'. The base classifiers are J48, RF, and the meta classifier is RF.
- Comparing the results of each method.

As mentioned earlier, NSL-KDD data set is used in the experiment. NSL-KDD is a data set which aimed to resolve some of the immanent problems of the KDD'99 data set [10]. Although NSL-KDD

data set suffers from some problems, it still can be applied as an effective benchmark data set to help us compare different anomaly detection methods.

The 20 per cent NSL-KDD training set is selected to train in the experiment. This 20 per cent data set is come from the web [11]. The original data set is only classified as normal and anomaly, which we call it Data Set_1. Furthermore, we also need to do some data processing to classify the instances as normal and the other 4 major attack classes.

We can use the method of mapping to add the feature of attack types to Data Set_1. According to the following Table 1 ,these attack types can be classified to four major classes DoS, U2R, R2L, Probe .We map 1 to DoS, 2 to U2R, 3 to R2L, 4 to Probe, 5 to normal. Then we can get a data set with a new class {DoS , U2R, R2L, Probe, normal},which we call it Data Set_ 2.

We can see from Table 1 that 32 different attack types are classified into 4 major classes. These four major attack classes are described as follows:

Denial of service (DoS) attacks: Attackers make hosts or network unavailable to its intended users by flooding the target machine or resource with superfluous requests. For example, ping-of-death, SYN flood.

Remote to Local (R2L) attacks: Attackers have access to local machine from a remote machine without authority. For example, "Warezmaster" attacks.

User to Root (U2R) attacks: Local attackers get highest authority (root) of local machine illegally. For example, "buffer overflow" attacks.

Probe: Attackers adopt programs to automatically scan networks in order to gather information or find vulnerabilities. For example port scanning and ping sweep.

Table 1. Attack Types

Class	Attack Types
DoS	Teardrop, Back, Udpstorm, Smurf, Neptune, Worm, Mailbomb, Pod, Processtable, Land, Apache2,.
U2R	Ps, Loadmodule, Perl, Rootkit, Xterm, Sqlattack, Buffer_overflow.
R2L	Sendmail, Named.Snmpguess,Warezmaster, Xlock, xsnoop, Imap, Phf, Snmpgetattack, Httptunnel,
Probe	Mscan, Ipsweep, Nmap , Saint ,Satan, Portsweep.

5. Experiments Results and Analysis

The experiment data set is divided into two parts. First, a 20 per cent NSL-KDD dataset is classified into normal and anomaly, which we call it Data set_1. Second, the feature of attack types is added to the 20 per cent NSL-KDD dataset, which we call it Data set_2. It is classified into normal and other four attacks classes such as DoS, U2R, R2L, Probe. 10 folds cross-validation method [12] is used both in Data set_1 and Data set_2. The results are compared on the basis of typical indexes and confusion matrix. The following six indexes are usually used to measure the performance of machine leaning methods. As we know, MCC and F-measure are very important features in the six indexes.

$$\text{Recall}=TP / (TP+FN) \quad (2)$$

$$\text{Precision}=TP / (TP+FP) \quad (3)$$

$$\text{Specificity}=TN / (TN+FP) \quad (4)$$

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (5)$$

$$\text{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP\times FP)+(TP\times FN)+(TN\times FP)+(TN\times FN)}} \quad (6)$$

$$\text{F-measure}=2\times(\text{Recall}\times\text{Precision}) / (\text{Recall}+\text{Precision}) \quad (7)$$

TP ---True Positives

FP --- False Positives
TN--- True Negatives
FN--- False negatives
MCC---Matthew's correlation coefficient
F-measure---the harmonic mean of Recall and Precision

5.1. Analysis the result of Data set_1

We can see the results of different methods which are used for Data set_1. They are displayed in Table 2.

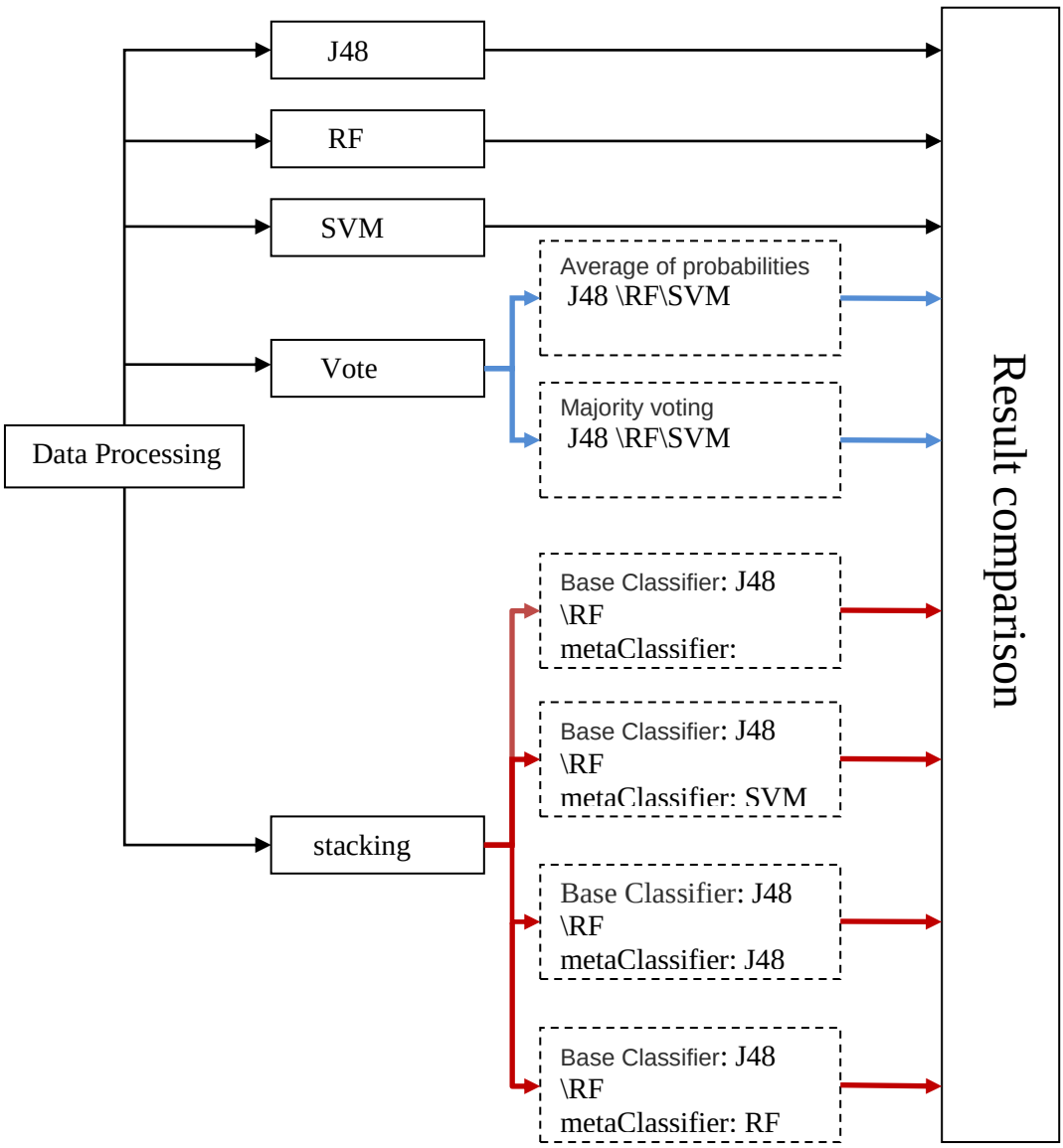


Figure 4. System Framework

Table 2. the experiment result of dataset_1

Method	CLASS	TP Rate	FP Rate	F-measure	MCC	Incorrectly Classified Instances	Correctly Classified Instances
J48	normal	0.996	0.004	0.996	0.991	111	25081
	anomaly	0.996	0.004	0.995	0.991		
RF	normal	0.996	0.006	0.995	0.990	124	25068
	anomaly	0.994	0.004	0.995	0.990		
SVM	normal	0.989	0.044	0.975	0.947	668	24524
	anomaly	0.956	0.011	0.971	0.947		
Vote-1 Average probabilities	normal	0.998	0.006	0.996	0.992	99	25093
	anomaly	0.994	0.002	0.996	0.992		
Vote-2 Majority voting	normal	0.999	0.004	0.997	0.994	71	25121
	anomaly	0.996	0.001	0.997	0.994		
Stacking-1 Meta: Logistic	normal	0.999	0.003	0.998	0.996	51	25141
	anomaly	0.997	0.001	0.998	0.996		
Stacking-2 Meta: SVM	normal	0.996	0.007	0.995	0.990	131	25061
	anomaly	0.993	0.004	0.994	0.990		
Stacking-3 Meta: J48	normal	0.996	0.006	0.995	0.990	129	25063
	anomaly	0.994	0.004	0.995	0.990		
Stacking-4 Meta: J48	normal	0.996	0.007	0.995	0.988	145	25047
	anomaly	0.993	0.004	0.994	0.988		

Table 3. the experiment result of dataset_2

Method	CLASS	TP Rate	FP Rate	F-measure	MCC	Incorrectly Classified Instances	Correctly Classified Instances
J48	DoS	1.000	0.000	1.000	1.000	4	25188
	U2R	0.909	0.000	0.952	0.953		
	R2L	0.990	0.000	0.995	0.995		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
RF	DoS	1.000	0.000	1.000	1.000	10	25182
	U2R	0.727	0.000	0.842	0.853		
	R2L	0.981	0.000	0.990	0.990		
	Probe	0.999	0.000	0.999	0.999		
	normal	1.000	0.001	1.000	0.999		
SVM	DoS	0.998	1.000	0.999	0.998	29	25163
	U2R	0.000	0.000	0.500	0.000		
	R2L	0.955	0.914	0.957	0.873		
	Probe	0.998	1.000	1.000	0.998		
	normal	1.000	1.000	1.000	1.000		
Vote-1 Average probabilities	DoS	1.000	0.000	1.000	1.000	4	25188
	U2R	0.818	0.000	0.900	0.904		
	R2L	0.990	0.000	0.995	0.995		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
Vote-2 Majority voting	DoS	1.000	0.000	1.000	1.000	5	25187
	U2R	0.818	0.000	0.900	0.904		
	R2L	0.990	0.000	0.995	0.995		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
Stacking-1 Meta: Logistic	DoS	1.000	0.000	1.000	1.000	6	25186
	U2R	0.909	0.000	0.833	0.836		
	R2L	0.990	0.000	0.993	0.993		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
Stacking-2 Meta: SVM	DoS	1.000	0.000	1.000	1.000	4	25188
	U2R	0.909	0.000	0.952	0.953		
	R2L	0.990	0.000	0.995	0.995		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
Stacking-3 Meta: J48	DoS	1.000	0.000	1.000	1.000	4	25188
	U2R	0.909	0.000	0.870	0.870		
	R2L	0.990	0.000	0.993	0.993		

Method	CLASS	TP Rate	FP Rate	F-measure	MCC	Incorrectly Classified Instances	Correctly Classified Instances
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		
Stacking-4 Meta: RF	DoS	1.000	0.000	1.000	1.000	2	25190
	U2R	0.909	0.000	0.952	0.953		
	R2L	1.000	0.000	0.995	0.995		
	Probe	1.000	0.000	1.000	1.000		
	normal	1.000	0.000	1.000	1.000		

In the three base classifiers, J48 has the high performance than SVM and RF. The F-measure is 0.996, MCC is 0.991.Vote-2 adopting the combinatorial rule of Majority Voting has a better performance than Vote-1 which has the rule of average probabilities. The F-measure is 0.997, MCC is 0.994. We can see that the method of Stacking-1 has the highest performance in the 9 methods. It uses J48, RF as two base classifiers, and uses logistic regression as the Meta classifier. The F-measure is 0.998, MCC is 0.996.

It can be seen from Table 2 that Vote-1 and Vote-2 both have better performance than the methods of J48, RF and SVM. It also can be seen that three stacking methods from Stacking-2 to Stacking-4 do not have better performance than the other two methods J48 and RF, although they are better than the single method SVM. But we can still draw a conclusion that stacking-1, which uses logistic regression as Meta classifier, is the best method to deal with Data set-1 in all the 9 methods. There are fewer incorrectly classified instances in stacking-1 method and highest F-measure and MCC as well.

5.2. Analysis the result of Data set_2

The results of different methods used for Data set_2 are displayed in table 3. The instances are classified into 5 classes.

In the three single classifiers, J48 still has the high performance. For the three classes of DoS, Probe and normal, F-measure and MCC are all 1.00. For U2R, F-measure is 0.952 and MCC is 0.953. For R2L, F-measure is 0.995 and MCC is 0.995

As to Vote method, here, Vote-2 adopting the combinatorial rule of Majority Voting has a similar performance with Vote-1 which has the combinatorial rule of average probabilities. The Incorrectly Classified Instances are 4 and 5 respectively.

In the four stacking methods, both Stacking-2 and Stacking-4 seem to have higher performance. Though its indexes of F-measure and MCC are the same, there are still some differences between them. We can see that TP rate is different. The former is 0.99 and the latter is 1.00. The difference can also be seen from the confusion matrix. The column of matrix means the prediction of classifier. The row of matrix means the real class of instance.

For method of stacking-2, from Table 4, it can be seen that there are 4 instances incorrectly classified. One instance of DoS is incorrectly classified to normal. One instance of U2R is incorrectly classified to normal. Another two instances of R2L are incorrectly classified to normal.

For method of stacking-4, from Table 5, it can be seen that only 2 instances are incorrectly classified. One instance of DoS is incorrectly classified to R2L and another instance of U2R is incorrectly classified to R2L.From the confusion matrix and typical indexes, we can draw a conclusion that stacking-4 using RF as Meta classifier has the best performance in the experiment of Data set_2.

Table 4. Confusion matrix of stacking-2

classified as	DoS	U2R	R2L	Probe	normal
DoS	9233	0	0	0	1
U2R	0	10	0	0	1
R2L	0	0	207	0	2
Probe	0	0	0	2289	0
normal	0	0	0	0	13449

Table 5. Confusion matrix of stacking-4

classified as	DoS	U2R	R2L	Probe	normal
DoS	9233	0	1	0	0
U2R	0	10	1	0	0
R2L	0	0	209	0	0
Probe	0	0	0	2289	0
normal	0	0	0	0	13449

6. Conclusion

In this paper, nine kinds of machine learning methods are used to deal with public network dataset NSL-KDD for anomaly detection. From the two part of the experiment, different ensemble learning technique can be evaluated. It is found that appropriate stacking method can get better performance than vote and other single classifiers. We also find that different ensemble learning method has different performance on different data set. For example stacking-1 using logistic regression as Meta classifier has the best performance on Data Set_1. And stacking -4 using RF as Meta classifier get the best performance on Data Set_2. In this experiment we can draw a conclusion that although different ensemble learning method has different performance, but we still can try to find an appropriate ensemble learning classifier which can achieve better performance than a single classifier for anomaly detection.

References

[1] T. Verwoerd,R. Hunt, “Intrusion detection techniques and approaches”, Computer Communications, vol. 25, 2002,pp.1356-1365, 2002.
[2] S. Revathi, A. Malathi, “A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection”, International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 12, pp1848-1853, 2013.
[3] Homoliak.I , Breitenbacher.D, Hanacek.P,”Convergence Optimization of Backpropagation Artificial Neural Network Used for Dichotomous Classification of Intrusion Detection Dataset”, Journal of computers. Vol. 2 Issue 22, pp.143-155, 2016.
[4] Aslahi-Shahri, B. M., Rahmani, R., Chizari, M., “ hybrid method consisting of GA and SVM for intrusion detection system”, Neural Computing Applications,Vol. 27,2016,pp.1669-1676 , 2016.
[5] Alotaibi, Bandar; Elleithy, Khaled,” A Majority Voting Technique for Wireless Intrusion Detection Systems”, 2016 IEEE Long Island Systems, Application And Technology Conference(LISAT),2016.
[6] J. R. Quinlan “Improved Use of Continuous Attributes in C4.5”, Journal of Articial Intelligence Research, issue 4,pp. 77-90, 1996
[7] Breiman. L, “ Random Forest”,Machine learning, Vol. 45 Issue 1, pp.5-32. 2001
[8] V.N. Vapnik, V. Vapnik, Statistical Learning Theory, vol. 2, Wiley, New York, 1998.
[9] http://www.cs.waikato.ac.nz/ml/weka/index.html
[10] https://web.archive.org/web/20150205070216/ http://nsl.cs.unb.ca/NSL-KDD/
[11] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
[12] Blum, Avrim, Adam Kalai, and John Langford. “Beating the hold-out: Bounds for k-fold and progressive cross-validation.” *Proceedings of the twelfth annual conference on Computational learning theory*. ACM, 1999.