2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)

Network intrusion detection based on random forest and support vector machine

Yaping Chang，WeiLi，Zhongming Yang
Computer Engineering Technical College
Guangdong Polytechnic of Science and Technology
Zhuhai, China
e-mail:llchangl@163.com

***Abstract*—The network intrusion detection techniques are important to prevent our system and network from malicious behaviors. In order to improve accuracy of network intrusion detection, machine learning, feature selection and optimization methods have been used, and the result tell us that the combination of machine learning and feature selection can improve accuracy. In this study, we developed a new machine learning approach for predicting network intrusion based on random forest and support vector machine. Since there were many potential features for network intrusion classification, random forest were used for feature selection based on variable importance score. We found that the host-based statistical features of network flow play an important role in predicting network intrusion. The performance of the support vector machine which used the 14 selected features on KDD 99 dataset has been evaluated by comparing it with the total(41) features and popular classifiers. The result showed that the selected features can achieve higher attack detection rate and it can be one of the competitive classifier for network intrusion detection.**

Keywords- network intrusion detection; random forest; support vector machine; feature selection

I. BACKGROUND

In our information society, the growth of internet application have been dynamically explosive. Thus we should pay more attention to network security in order to protect the information from network attacks and threats. The network intrusion detection techniques are important to prevent our system and network from malicious behaviors. Network intrusion detection is to determine the current network behavior intrusion or normal base on the collection data of network status which high dimension and linear inseparable. In order to improve accuracy of network intrusion detection, many researchers use machine learning, feature selection and optimization methods, such as partial least squares (PLS) and kernel vector machine[1], Autoencoder Network with Feature Reduction[2], Ant Colony Optimization algorithm and Support Vector Machine(ACO-SVM)[3], simplified swarm optimization(SSO)[4], FAST feature selection algorithm, Adaptive Binary Quantum-Inspired Gravitational Search Algorithm and Support Vector Machine (FAST-ABQGSA-SVM)[5], Fisher feature selection and K-nearest neighbor algorithm[6], Naive Bayes classifier[7], Transductive confidence machines for k-nearest neighbors (TCM-KNN)[8] and so on. The result tell us that the combination of machine learning and feature selection can improve accuracy of network intrusion detection. In this study, we use random forest to select the important feature and SVM as the classification algorithm.

II. MATERIALS AND METHODS

A. Random forest for feature selection

In this study, important features for network intrusion detection were identified by using the random forest(RF) algorithm. The RF algorithm uses a combination of independent decision trees to model data and measure variable importance[9]. Each decision tree in the forest is constructed using a bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree. These instances are called the out-of-bag(oob) data for the tree. In every tree grown in the forest, put down the oob cases and count the number of votes cast for the correct class. Now randomly permute the values of variable m in the oob cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-m-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable m. For a fixed number of trees in the forest, the larger importance score a variable has, the more important it is for classification. In addition, a Z-score can be obtained by dividing the variable importance score by its standard error, and a statistical significance level may be assigned to the z-score assuming normality[9]. The RF algorithm can handle many redundant features and avoid model overfitting. It has been shown that RF outperform AdaBoost ensembles on noisy datasets, and can perform well on data with many weak input variables.

B. Support vector machine classifier

SVM is developed on the principle of structural risk minimization. It is one of the machine learning algorithm that map the training data into the high-dimensional feature space through some nonlinear mapping. In this study, support vector

This study was supported by Guangdong Province outstanding young teacher training plan (Yq2014187) and the Higher vocational education in Guangdong Province(2016gzpp007).

978-1-5386-3221-5/17 $31.00 © 2017 IEEE
DOI 10.1109/CSE-EUC.2017.118

machine(SVM) were trained with the selected features to predict network intrusion. The svm software package(available at http://www.csie.ntu.edu.tw/~cjlin/libsvm) was used to construct svm classifiers. SVM offer the following advantages over conventional statistical learning algorithms: 1) High generalization performance even with high dimension feature vectors; 2) The ability to manage kernel functions that map input data to higher dimensional space without increasing computational complexity[10].

The main idea of SVM is to create an optimal hyperplane to classify the data into two classes(positive and negative) and maximize distance between the hyperplane separating the two classes and the closet data points to the hyperplane. The optimal hyperplane maximizes the separation margin between the two classes of training data, and is defined by a fraction of the input data instances(Called support vectors) close to the hyperplane. The distance measurement between the data points in the high-dimensional space is defined by the kernel function. In this study, we used the radial basis function(RBF) kernel (1):

$$\mathrm{k}(x, y) = \exp\left(-\gamma \|x - y\|^2\right) \quad (1)$$

Where x and y are two data vectors, and γ is a training parameter. A smaller γ value makes the decision boundary smoother. Another parameter for SVM training is the regularization factor c , which controls the trade-off between low training error and large margin.

C. Data

In this study we use the KDD 99 dataset as the dataset of the experiment.This dataset can be divided into five categories which are Normal, DOS, R2L, U2R and Probing. Each network record contains 41 features shown in table Ⅰ , of which 32 attributes are continuous and 9 attributes are discrete. In order to compared with Li Cong[5], we randomly selected 29572 samples form the 10 percent training set as our training set. Because there are numerous Normal and DOS samples, thus chose from Normal and DOS samples. Test set include 6000 samples, the number of samples shown in table Ⅱ.

TABLE I. THE 41 FEATURES OF KDD 99 DATASET

Number	***Feature name***	***Data type***
1	duration	continuous
2	protocol_type	discrete
3	service	discrete
4	flag	discrete
5	src_bytes	continuous
6	dst_bytes	continuous
7	land	discrete
8	wrong_fragment	continuous
9	urgent	continuous
10	hot	continuous
11	num_failed_logins	continuous
12	logged_in	discrete
13	num_compromised	continuous
14	root_shell	discrete
15	su_attempted	discrete
16	num_root	continuous
17	num_file_creations	continuous
18	num_shells	continuous
19	num_access_files	continuous
20	num_outbound_cmds	continuous
21	is_hot_login	discrete
22	is_guest_login	discrete
23	count	continuous
24	srv_count	continuous
25	serror_rate	continuous
26	srv_serror_rate	continuous
27	rerror_rate	continuous
28	srv_rerror_rate	continuous
29	same_srv_rate	continuous
30	diff_srv_rate	continuous
31	srv_diff_host_rate	continuous
32	dst_host_count	continuous
33	dst_host_srv_count	continuous
34	dst_host_same_srv_rate	continuous
35	dst_host_diff_srv_rate	continuous
36	dst_host_same_src_port_rate	continuous
37	dst_host_srv_diff_host_rate	continuous
38	dst_host_serror_rate	continuous
39	dst_host_srv_serror_rate	continuous
40	dst_host_rerror_rate	continuous
41	dst_host_srv_rerror_rate	continuous

TABLE II. NUMBER OF SAMPLES

Type	***Number of samples in training set***	***Number of samples in test set***
Normal	14000	2900
DOS	10287	1200
U2R	52	20
Probing	4107	806
R2L	1126	1074
Total	29572	6000

D. Data Preprocessing

1) Coding the char feature

There are three char features in 41 features, namely protocol_type、service and flag, then coding them before train and test. In this study, real numeric values are used to represent char feature, such as, 1 for tcp, 2 for udp, 3 for icmp in protocol_type; In flag type 1 for oth, 2 for rej, 3 for rsto, 4 for rstos0, 5 for rstr, 6 for s0, 7 for s1, 8 for s2 and so on; In Serivce type 1 for aol, 2 for auth, 3 for bgp, 4 for courier , and so on. After coding each sample is represented by 41 features vector and 1 category.

2) Min-Max normalization

Because the range of feature values is very large, and svm determine the sample category is based on the distance between sample and hyperplane, thus a few number of features have much influence on svm performance, others have little influence. In order to eliminate this adverse effect, we do min-max normalization before train and test, after scaling every feature is between 0 and 1. The min-max normalization formula is (2),

$$x_i^{'} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \qquad (2)$$

$\mathrm{x}_i^{'}$ is the ith feature value after min-max normalization, x_i is the ith feature value before min-max normalization, $\min(x_i)$ is the minimum of ith feature, $\max(x_i)$ is the maximum of the ith feature.

3) Classifier performance evaluation

The following measurements are often used to evaluate the efficiency of the classifier:

$$\text{detection rate=}\frac{\mathrm{TN}}{TN+FP} \qquad (3)$$

$$\text{False alarm rate=}\frac{\mathrm{FN}}{\mathrm{FN+TP}} \qquad (4)$$

In (3) and (4), the TP represents that the normal sample is correctly predicted, FP indicates that the abnormal sample is predicted as normal, FN denotes that the normal sample is wrongly predicted as abnormal, and TN represents the abnormal behavior is correctly detected.

III. Result and discussion

There are many potential features for network intrusion detection. To select the important features, samples were coded with the 41 features, and then used to construct random forest. There are 500 trees in the forest, and mtry is 6(we use the default parameters), selected the top 14 features based on variable importance shown in Figure 1.

The important features selected by Random forest are 2, 3, 5, 6, 12, 23, 24, 32, 33, 34, 35, 36, 37 and 40, which shown in table Ⅲ . From the result we found that the host-based statistical features of network flow play an important role in network intrusion detection. In 14 selected important features 7 are the host-based statistical features of network flow.

Dr. Saurabh Mukherjee also found that the 3, 5, 6, 12, 23, 24, 32, 33, 36 and 40 are important features in predicting network intrusion[7]. Yinhui Li[11] also found that the 2, 32, 33, 34, 35, 36, 37 and 40 are important features, but they have not found other important features. Yinhui Li[11] also found that the host-based statistical features of network flow play an important role in predicting network intrusion.

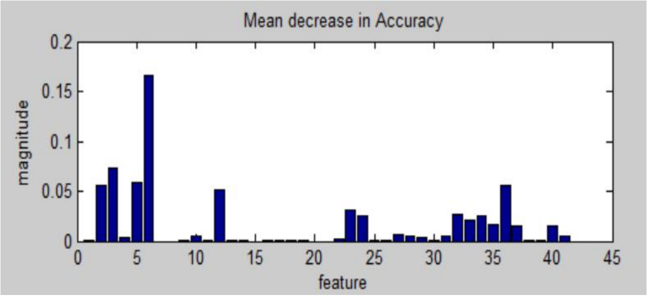

Figure 1. Variable importance

TABLE III. Important features selected by random Forest

number	*Feature name*	*category*
2	protocol_type	basic features of TCP connection
3	service	
5	src_bytes	
6	dst_bytes	
12	logged_in	content features of TCP connection
23	count	The time-based statistical features of network flow
24	srv_count	
32	dst_host_count	The host-based statistical features of network flow
33	dst_host_srv_count	
34	dst_host_same_srv_rate	
35	dst_host_diff_srv_rate	
36	dst_host_same_src_port_rate	
37	dst_host_srv_diff_host_rate	
40	dst_host_rerror_rate	

TABLE IV. Performance of support Vector machine

feature	*Attack detection rate*	*False alarm rate*
Selected 14 features	93%	3%
total(41) features	90%	2%
RS-GA-SVM	88.2%	2%

Table Ⅳ shown the performance of support vector machine using 41 features and the selected 14 features(using default parameter). When use 14 features the attack detection rate can higher than total(41) features, and higher than the Rough set genetic algorithm support vector machine(RS-GA-SVM) method[5]. Furthermore, the false alarm rate of 14 selected features is slightly higher than that of total feature SVM classifier while the training and predicting time is greatly reduced.

IV. CONCLUSION

We have developed a new machine learning approach for predicting network intrusion based on random forest and support vector machine. Since there were many potential features for network intrusion classification, random forest were used for feature selection based on variable importance score. The important features selected by Random forest are 2, 3, 5, 6, 12, 23, 24, 32, 33, 34, 35, 36, 37 and 40. The host-based statistical features of network flow play an important role in predicting network intrusion. Only use 14 features can get higher Attack detection rate than total(41) features.

For future work, we want to develop more model to predict network intrusion and research some other optimization algorithm to optimize svm parameters.

[1] Liyun Wu, shenglin Li, Xusheng Gan, Minghua Wang, “Network anomaly intrusion detection CVM model based on PLS feature extraction,” Control and Decision. China, vol. 32, pp.755-758, 2017.

[2] Ni Gao, Ling Gao , Yiyue He, Hai Wang, “A Lightweight intrusion detection model based on autoencoder network with feature reduction,” ACTA ELECTRONICA SINICA. China, vol. 45, pp.730-739, 2017.

[3] Guorong Xiao, “Network intrusion detection by combination of improved ACO and SVM,” Computer Engineering and Applications. Guangzhou, vol. 50, pp. 75–78, 2014.

[4] Yuk Ying Chung, Noorhaniza Wahid, “A hybrid network intrusion detection system using simplified swarm optimization,” Applied Soft Computing. Australia, vol. 12, pp. 3014-3022, 2012.

[5] Cong Li, Renwu Yan, Changshui Zhu, Guangyin Gao, “Network intrusion detection based on Fast feature selection and ABQGSA-SVM,” Application Research of Computers. China, vol. 33, pp. 75–78, 2017.

[6] Yafen Cui, Nannan Xie, “An iutrusion detection method based on Feature selection,” Journal of Jilin University(Science Edition). China, vol. 53, pp. 112–116, 2015.

[7] Dr. Saurabh Mukherjee, Neelam Sharma, “Intrusion Detection using Naive Bayes Classifier with Feature Reduction,” Procedia Technology. India, vol. 4, pp. 119-128, 2012.

[8] Yang Li, Li Guo, “An active learning based TCM-KNN algorithm for supervised network intrusion detection,” Computers & Security. china, vol. 26, pp. 459-467, 2007.

[9] Breiman L, “Random forests,” Machine Learning. berkeley, vol. 45, pp. 5-32, 2001.

[10] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: a library for support vector machines,”ACM Transactions on Intelligent Systems and Technology. Taiwan, vol.2, pp.1-27, 2011 .

[11] Yinhui Li, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, Kuobin Dai, “An efficient intrusion detecion system based on support vector machines and gradually feature removal method,” Exper Systems with Applications. China, vol.39, pp.424-430, 2012.