Construction and Application of Machine Learning Model in Network Intrusion Detection

Construction and Application of Machine Learning Model in Network Intrusion Detection

Guanglei Qi[1*], Zhijiang Chen[1], Haiying Zhao[1,2], Chensheng Wu[1]

[1]Mobile Media and Cultural Computing Key Laboratory of Beijing, Century College, Beijing University of Posts and Telecommunications, Beijing, 102101, China

[2] Beijing University of Posts and Telecommunications, Beijing, 100876

* qgl@emails.bjut.edu.cn

Abstract: The modeling of network intrusion detection is a kind of significant network security protection technology. Nowadays, the network intrusion detection model can not accurately describe the intrusion behavior, so that incomplete network intrusion detection will occur. Therefore, based on machine learning algorithm, a network intrusion detection model is designed in the paper, which supports the fact that vector machine (SVM) fits the mapping relationship between network intrusion detection characteristics and network intrusion behavior. Meanwhile, a network intrusion detection model that reflects the relationship between the two aspects is established in the paper too. As a result, the experimental results show that the model proposed in the paper can not only accurately identify the network intrusion behavior, but has a very fast detection speed. Moreover, the model proposed in the paper has obtained much better network intrusion detection results than that of other models, which is of wide application prospect.

1. Introduction

In cyberspace security, network security plays an important role, and the security of network infrastructure provides the basis for the reliable operation of the Internet. Although various network security detection measures are designed for secure communications in the development of various Internet activities, machine learning technology has a wide range of applications in plenty of areas, such as BGP anomaly detection, malicious domain name detection, botnet detection, network intrusion detection, and malicious encrypted traffic identification.

Domain name system is one of the core applications in the Internet, which often becomes the target of attack, or is used as an attack tool by attackers. Therefore, for a long time, the security of the domain name system has been the research focus in the network security. Additionally, the early malicious domain name detection method aims to set a malicious domain name blacklist or an interception list in the domain name system, firewall or network intrusion detection system, which can be circumvented easily by attackers. Then, the following method based on inquiry number has many problems, such as high mistake rate and no ability to detect unknown abnormal domain name. In recent years, the machine learning technology is aplied to construct detection rules for malicious domain names, which has been a new research direction in this field[1-2].

Thr intrusion detection is a reasonable complement for the firewall to assist the system to tackle network attacks and expand administrator's security management capabilities in the system, including security auditing, monitoring, attack identification and response, so that the integrity of the

information security infrastructure can be improved.

In a successful intrusion detection system, not only system administrators are allowed to keep abreast of any changes in network systems, including programs, files, and hardware devices, etc., but guidelines are provided for the development of network security policies. Moreover, the system should be simple enough to be managed and configured, so that the network security can be easily obtained by non-professionals. What is more, the scale of intrusion detection can be changed only according to the changes in cyber threats, system architecture, and security requirements. Then, the intrusion detection system responds immediately after discovering the intrusion, including cutting off the network connection and recording events and alarms.

2. Malicious domain name detection process based on machine learning

Based on machine learning, malicious domain name detection is usually implemented by an offline model and an online model. The general process is shown in Figure 1.

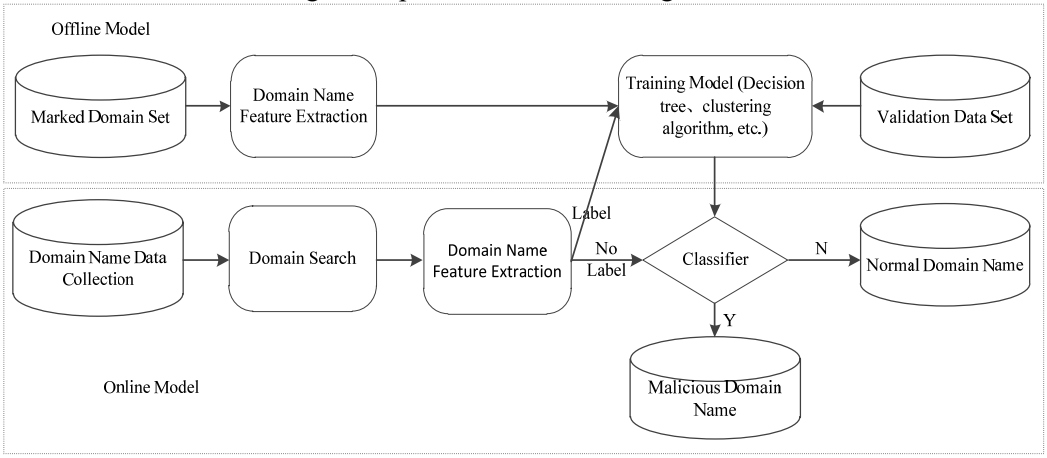


Fig 1. Malicious domain name detection process based on machine learning

In the offline model, the tagged legitimate domain name and malicious domain name are adopted as training data sets to extract features according to the network layer, regions, time, DNS responses, TTL, or domain name information. Then, decision tree and X-Means clustering algorithm are selected to construct the training model as well as some known domain name data sets which are provided by sites, such as malwareurl.com, McAffee Site Advisor, or Norton Safe Web, so that the training models can be verified and adjusted. What is more, in the online detection model, the domain name traffic that is collected timely in the network is subjected to passive domain name query analysis and domain name feature extraction. If it is a known domain name information, that is, the marked domain name feature, the training model will be input to continue the following training, while if it is an unknown domain name information, in other words, if it is a feature without tags, a trained classifier will be input to determine whether the domain name is a malicious domain name or not.

3. Network intrusion detection

3.1 Research Status

Nowadays, the research on the network intrusion detection has been deepened, and plenty of network intrusion detection models with better performance have emerged. Meanwhile, the current network intrusion detection models can be roughly divided into two categories, namely misuse detection and anomaly detection. Misuse detection is the most primitive intrusion detection technology, which constructs a database of network intrusion detection to match the to-be-detected behavior with intrusion behavior in the database. If it matches, it will be classified into the corresponding intrusion category. Otherwise, it will be a normal behavior[3-4].

In practical applications, the misuse detection model can only detect existing intrusions. In other words, new intrusions can not be detected by the misuse detection model. Therefore, when there are new intrusions, the model will be powerless, and the actual application value wil be low. In addition, compared with misuse detection technology, anomaly detection technology belongs to pattern recognition. With the help of analyzing intrusion behaviors through certain rules, some new and never-present intrusion behaviors can be detected. Moreover, since the practical application value is high, it has become an important direction of current network security and field research.

In the anomaly detection process of network intrusion, the classifier selection of intrusion behavior is very critical. Nowadays, neural network for constructing network intrusion behavior classifiers is the commonest method, and neural network is a modeling method based on big data theory, which requires training samples. Therefore, the cost of network intrusion detection increases. Meanwhile, since the sample of network intrusion is really few, it is difficult to meet the requirements of many samples. The network intrusion detection results of the neural network are not stable. Sometimes, the detection accuracy is high. Sometimes, it is low[5].

In recent years, with the continuous deepening on machine learning theory, a new type of modeling technology, support vector machine (SVM), has been created. Compared with neural network, the support vector machine (SVM) has been widely applied in practice. Meanwhile, the number of training samples is not so high, and the learning performance is not better than that of neural networks. Therefore, some scholars have introduced it into the application of network intrusion detection.

In the process of network intrusion detection modeling based on support vector machine, there are two problems. One is the determination of the parameters of support vector machine. Some scholars use gradient descent algorithm and genetic algorithm to obtain the optimization. However, optimization looking time of the gradient descent algorithm is long, which will affect the efficiency of network intrusion detection. The other is that there is no unified theoretical guidance for genetic operator setting of genetic algorithm. Therefore, it is easy to obtain the local optimal parameter value, affecting the network intrusion detection results.

In order to obtain high-accuracy network intrusion detection results, a network intrusion detection model based on the ant colony algorithm to determine the parameters of the support vector machine is proposed in the paper according to the limitations of the current network intrusion detection model. Meanwhile, one-to-many network intrusion detection classifier is established by SVM, and ant colony algorithm is used to determine the optimal parameters. Additionally, the current standard network intrusion detection database is applied to test the validity of the model, and the accuracy is more than 95%. The detection error is far lower than the actual application range.

3.2 Support Vector Machine

Support Vector Machine (SVM), proposed by Vapnik et al., is a kind of machine learning algorithm with excellent performance, which differs from the working principle of neural networks. It is modeled based on the principle of minimization of structural risk. In addition, it is a two-category algorithm. An optimal plane is found to divide all training samples into two categories. One is above the plane and the other is below the plane. Meanwhile, keep the sample as far away from the optimal plane as possible. The sample that is above the optimal plane is called support vector, whose working principle is shown in Figure 2.

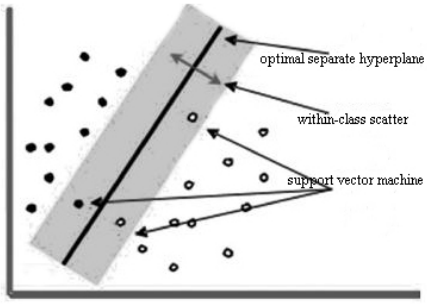


Fig 2. Schematic diagram of optimal classification plane

For the set $\{(x_1, y_1),(x_1, y_1),....,(x_i, y_i),....,(x_n, y_n)\}$ with n samples, the function $\varphi(x)$ is applied to map the samples. Moreover, the sample is classified in the mapping space. There is

$$f(x) = \mathrm{sgn}(w \cdot \varphi(x) + b) \tag{1}$$

where w is the weight, and b is the threshold.

To find the optimal separate hyper plane, the optimal w and b values must be found. It is very difficult to solvie the formula (1) directly to obtain the optimal w and b values. Therefore, the following constraints based on the structural risk minimization principle should be set.

$$y_i \cdot (w \cdot \varphi(x_i) + b) \geq 1 \tag{2}$$

To speed up the modeling process, a slack variable A is adopted to perform a trade off between the classification accuracy and the classification error, so that the optimal separate hyper plane can be transformed into the following form.

$$\min \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \xi_i \tag{3}$$

The corresponding constraint is as follows.

$$y_i (w \cdot x_i) + b) \geq 1 - \xi_i \qquad \xi_i \geq 0, i = 1, 2, ..., n \tag{4}$$

where C refers to the penalty of the error.

Introducing L multiplier A to get the dual form of formula 4,

$$\min \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\varphi(x_i) \cdot \varphi(x_j)) + \sum_{i=1}^{n} \alpha_i \tag{5}$$

there will be the following constraint:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C \tag{6}$$

For the nonlinear classification problem, the kernel function K should be introduced to get

$$\min \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{n} \alpha_i \tag{7}$$

where $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$

In addition, the optimal separate hyper plane for support vector machines is as follows.

$$f(x) = \mathrm{sgn}(\sum_{i,j=1}^{n} \alpha_i y_i k(x_i, x) + b) \tag{I(8)}$$

Selecte the radial basis function,

$$k(x, x_j) = \exp(-\frac{\| x - x_j \|}{2\sigma^2}) \tag{9}$$

where σ refers to the kernel width parameter.

3.3 Impact of Parameters on Network Intrusion Detection

Analyzing the working principle of SVM, it can be found that the influence of parameters C and σ on their own learning performance is of great signifivance. What is more, a training sample is selected to analyze the correct rate of the network intrusion detection under different parameters. Table 1 shows the results. Analyzing Table 1, it is found that even if the environment and data are the same, the difference in the accuracy of the intrusion detection for different parameters is still large. Therefore, the optimal values of parameters C and σ need to be selected.

Table 1. Influence of parameter C and σ on support vector machine learning performance

C	σ	Intrusion detection accuracy /%
10	0.01	61.85
50	0.1	98.67
100	1	73.02
500	10	77.89
1000	100	96.12
5000	1000	66.97
10000	2000	78.54

4. Construction of Network Intrusion Model

In network intrusion detection, LSSVM parameter optimization problem can be expressed according to the following formula:

$$\max P(C,\sigma) \quad s.t. \begin{cases} C \in [C_{\min}, C_{\max}] \\ \sigma \in [\sigma_{\min}, \sigma_{\max}] \end{cases} \tag{10}$$

The steps of the network intrusion detection are as follows:

Step1: Network status information is collected to extract features of network intrusion detection and perform the following processing on features:

$$x_1 = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{11}$$

where $x_{\max}$ and $x_{\min}$ refers to the maximum and minimum values respectively.

Step2: The SVM parameters (C,σ) are considered as a path of ant colony crawling, and the network intrusion detection training samples are modeled according to each set of parameters, so that different detection accuracy rates can be obtained.

Step3: Through the pheromone update operation and node transfer of the ant colony, the path will be crawled. The optimal parameter (C,σ) combination is eventually found with the help of the path optimization.

Step4: An optimal network intrusion detection model is established based on the optimal parameters (C,σ) combination.

Due to the classification problem of support vector machines for two categories, there are many kinds of network intrusion behaviors, such as denial of service attacks, unauthorized remote access attacks, and port scanning attacks. Therefore, a one-to-one approach is used in the paper to build multiple classifiers, which is as shown in Figure 3.

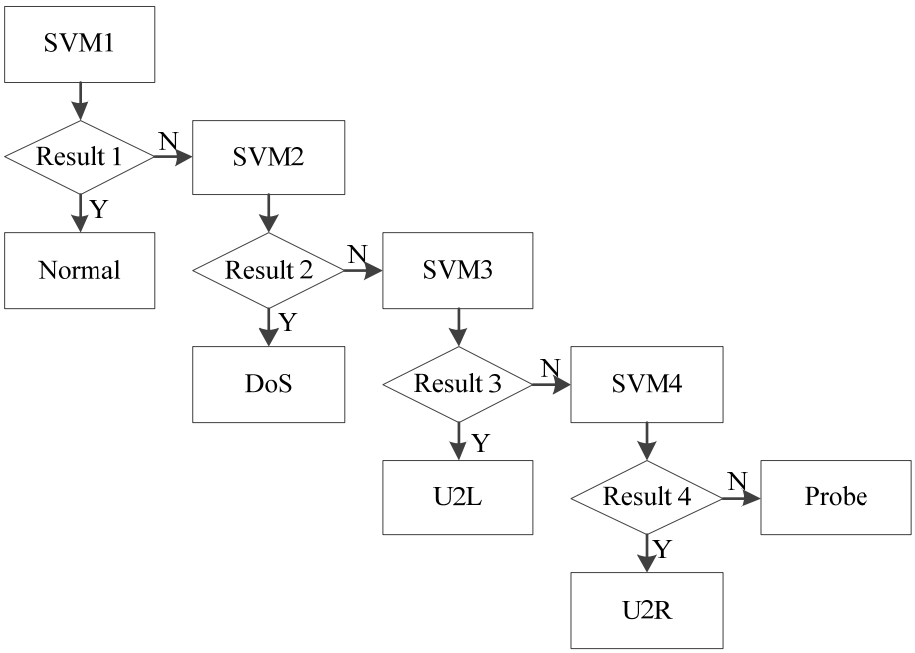


Fig. 3 Classifier structure of network intrusion detection

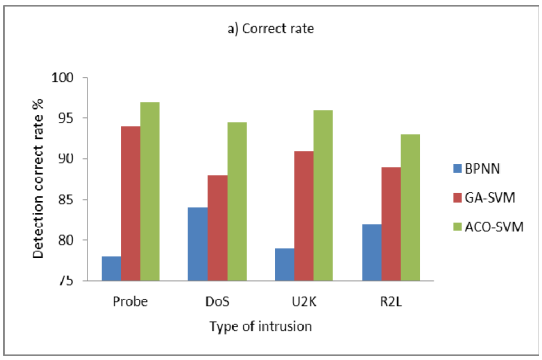
5. Experimental results and analysis

A network intrusion detection data set is selected as a test object, including four network intrusion behaviors, that is, Do S, Probe, U2R, and R2L. In addition, since the data set is very large, 10% of the data is selected for specific experiments. Therefore, in order to make the experimental results convincing, BP neural network (BPNN) and genetic algorithm are applied to optimize the network intrusion detection model of SVM (GA-SVM).

As the comparison model, the following indicators are used as the evaluation criteria for the experimental results.

Correct rate = number of correct detection of samples / total number of samples × 100%

The simulation results are shown in Figure 4. From Figure 4, it can be seen that among all models, the ACO - SVM has the highest network intrusion detection accuracy, which is followed by GA-SVM. The lowest accuracy of network intrusion detection is BPNN, which has the lowest false positive rate at the same time, indicating that the ACO-SVM can accurately recognize network intrusion behavior and get the ideal detection result. Moreover, it can be obtained from Figure 4b) that ACO-SVM network intrusion detection takes a minimum time, meeting the efficiency requirements of network intrusion detection, and the superiority of network intrusion detection result is very obvious.



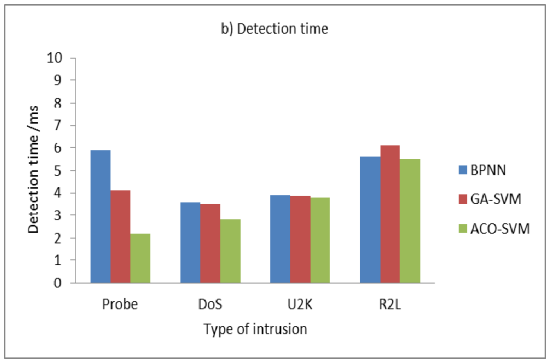


Fig 4. Comparison of network intrusion detection results

6. Conclusion

Network intrusion detection in network security research, including malicious domain name modification and other network intrusion situations, is analyzed in the paper. Meanwhile, SVM is applied to construct the model and verify the model. Therefore, compared with the traditional network intrusion model, the advantages of the model proposed in the paper has been fully reflected.

References:

[1] Wang G, Konolige T, Wilson C, et al. You are how you click: clickstream analysis for Sybil detection//Proceedings of the 22rd USENIX Security Symposium. Washington, USA, 2013: 241-256

[2] Freeman DM. Using naive bayes to detect spammy names in social networks// Proceedings of the ACM Workshop on Security and Artificial Intelligence. Berlin, Germany, 2013: 3–12

[3] Viswanath B, Bashir M A, Crovela M, et al. Towards detecting anomalous user behavior in online social networks//Proceedings of the 23rd USENIX Security Symposium. San Diego, USA, 2014:223-238

[4] Egele M, Stringhini G, Kruegel C, et al. Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable & Secure Computing, 2015, 12(2): 91-98

[5] Thomas K,Grier C,Ma J,et al. Design and evaluation of a real-time url spam filtering service//Proceedings of the Symposium on Security and Privacy. Oakland, USA, 2011: 447-462