Atlantis Highlights in Engineering, volume 5
International Conference on Precision Machining, Non-Traditional Machining and Intelligent Manufacturing (PNTIM 2019)

Construction and Application of Machine Learning Model in Network Intrusion Detection

Guanglei Qi*
1Mobile Media and Cultural Computing Key Laboratory of Beijing, Century College
Beijing University of Posts and Telecommunications
Beijing, 102101, China
E-mail: qgl@emails.bjut.edu.cn

Haiying Zhao[1,2]
[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

Zhijiang Chen
Mobile Media and Cultural Computing Key Laboratory of Beijing, Century College
Beijing University of Posts and Telecommunications
Beijing, 102101, China

Chensheng Wu
Mobile Media and Cultural Computing Key Laboratory of Beijing, Century College
Beijing University of Posts and Telecommunications,
Beijing, 102101, China

***Abstract*—The modeling of network intrusion detection is an important network security protection technology. The current network intrusion detection model can not accurately describe the intrusion behavior, resulting in incomplete network intrusion detection. Therefore, a network intrusion detection model based on machine learning algorithm was designed. In addition, support vector machine (SVM) fits the mapping relationship between network intrusion detection characteristics and network intrusion behavior, and established a network intrusion detection model that reflected the relationship between the two aspects. Finally, the experimental results showed that the model not only can accurately identify the network intrusion behavior, but also has a very fast detection speed. It has obtained better network intrusion detection results than other models, and had a wide application prospect.**

Keywords-Network Intrusion Detection; Machine Learning; SVM; Network Security

I. INTRODUCTION

Network security is an important aspect in cyberspace security. The security of network infrastructure provides the basis for the reliable operation of the Internet. And various network security detection measures provide secure communications for the development of various Internet activities. Machine learning technology has a wide range of applications in 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. It often becomes the target of attack, or is used by attackers as an attack tool. Therefore, the security of the domain name system has always been the research focus of network security. The early malicious domain name detection method was to set a malicious domain name blacklist or an interception list in the domain name system, firewall or network intrusion detection system. This method can easily be evaded detection by attackers. The followed method based on inquiry number has problems of high

mistake rate and no ability to detect unknown abnormal domain name. In recent years, application of machine learning technology to construct detection rules for malicious domain names is a new research direction in this field[1-2].

Intrusion detection is a reasonable complement to the firewall to help the system deal with network attacks, expand the system administrator's security management capabilities, including security auditing, monitoring, attack identification and response, and improve the integrity of the information security infrastructure.

For a successful intrusion detection system, it not only allows system administrators to keep abreast of any changes in network systems (including programs, files, and hardware devices, etc.), but also provides guidelines for the development of network security policies. More importantly, it should be simple to manage and configure, so that non-professionals can easily obtain network security. Moreover, the scale of intrusion detection should also change based on changes in cyber threats, system architecture, and security requirements. The intrusion detection system responds promptly after discovering the intrusion, including cutting off the network connection, recording events and alarms.

II. MALICIOUS DOMAIN NAME DETECTION PROCESS BASED ON MACHINE LEARNING

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

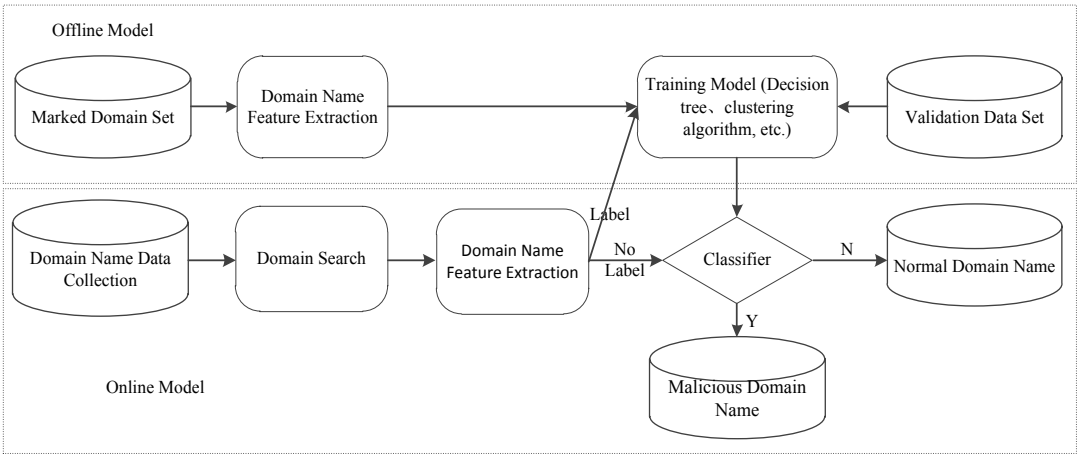

Figure 1. Malicious domain name detection process based on machine learning

In the offline model, the tagged legitimate domain name and malicious domain name are used as training data sets to extract features based on the network layer, regions, time, DNS responses, TTL, or domain name information. Then decision tree and X-Means clustering algorithm are selected to build the training model, and some known domain name data sets provided by sites such as malwareurl.com, McAffee Site Advisor, or Norton Safe Web to verify and adjust training models. In the online detection model, the domain name traffic collected in real time in the network is subjected to passive domain name query analysis and domain name feature extraction. If it is a known domain name information, namely, the marked domain name feature, the training model is input to continue training, while if it is an unknown domain name information, namely, the feature without tags, then a trained classifier is input to determine whether the domain name is a malicious domain name.

III. NETWORK INTRUSION DETECTION

A. Research Status

At present, the research on network intrusion detection has been deepened, and many network intrusion detection models with better performance have emerged. The current network intrusion detection models can be roughly divided into two categories: misuse detection and anomaly detection. Misuse detection is the most primitive intrusion detection

technology. It constructs a database of network intrusion detection, and matches 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 is a normal behavior[3-4].

In practical applications, the misuse detection model can only detect existing intrusions and can't detect new intrusions. Therefore, when there are new intrusions, the model is powerless, and the actual application value is low. Compared with misuse detection technology, anomaly detection technology belongs to pattern recognition. Through analysis of intrusion behaviors through certain rules, some new and never-present intrusion behaviors can be detected. The actual application value is relatively high and becomes the current network security and an important direction of field research.

In the anomaly detection process of network intrusion, classifier selection of intrusion behavior is very critical. At present, neural network for the construction of network intrusion behavior classifiers is the commonest method. Neural network is a modeling method based on big data theory and requires training samples, so the cost of network intrusion detection is increased. At the same time, the sample of network intrusion is really few and it is difficult to meet the requirements of many samples. Therefore, the network intrusion detection results of the neural network are not stable, and the detection accuracy is sometimes high or sometimes low[5].

In recent years, with the continuous deepening of machine learning theory, a new type of modeling technology has been created, that is, support vector machine (SVM). Compared with neural network, and support vector machine (SVM) has been used. The number of training samples is not so high, and the learning performance is not better than that of neural networks. For this reason, 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 the following two problems. First is determination of the parameters of support vector machine. For the problem of parameter determination, some scholars use the gradient descent algorithm and the genetic algorithm to obtain the optimization, however, optimization looking time of gradient descent algorithm is long, which will affect the efficiency of network intrusion detection. Second, there is no unified theoretical guidance for genetic operator setting of genetic algorithm, so it is easy to obtain the local optimal parameter value and affect the network intrusion detection results.

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

B. Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm with excellent performance and proposed by Vapnik et al. It differs from the working principle of neural networks. Actually, it is modeled based on the principle of minimization of structural risk and is a two-category algorithm. Find an optimal plane and divide all training samples into two categories: one above the plane and the other below the plane. At the same time, keep the sample as far away from the optimal plane as possible. The sample above the optimal plane is called support vector and its working principle is shown in Figure 2.

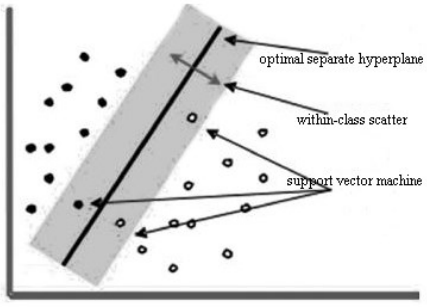

Figure 2. Schematic diagram of optimal classification plane

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 used to map the samples, and then the sample is classified in the mapping space, then there is

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

where w is the weight; b is the threshold.

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

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

In order to speed up the modeling process, a slack variable A is used 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

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

where C represents the penalty of the error.

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

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

and there is 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, we should introduce the kernel function K, so we can 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 $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$

The optimal separate hyper plane for support vector machines is

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

Selecting the radial basis function that is

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

where σ represents the kernel width parameter.

C. Impact of Parameters on Network Intrusion Detection

Analyzing the working principle of SVM, we can find that the influence of parameters C and σ on their learning performance is very important. A training sample is selected to analyze the correct rate of network intrusion detection under different parameters, and Table 1 shows the results. Through analysis of Table 1, we find that even if the environment and data are the same, the difference in accuracy of intrusion detection for different parameters is still large. Therefore, the optimal values of parameters C and σ need to be selected.

TABLE I. 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

IV. CONSTRUCTION OF NETWORK INTRUSION MODEL

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

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

The steps of network intrusion detection are as follows:

Step1: Collect network status information, extract features of network intrusion detection, and perform the following processing on features:

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

Where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values respectively.

Step2: The SVM parameters (C,σ) are regarded as a path of ant colony crawling. The network intrusion detection training samples are modeled according to each set of parameters to obtain different detection accuracy rates.

Step3: Through the pheromone update operation and node transfer of the ant colony, the path is crawled. Finally, the optimal parameter (C,σ) combination is found through path optimization.

Step4: Establish an optimal network intrusion detection model based on the optimal parameters (C,σ) combination.

Because of the classification problem of support vector machines for two categories, there are many types of network intrusion behaviors, such as denial of service attacks, unauthorized remote access attacks, and port scanning attacks. This article uses a one-to-one approach to build multiple classifiers, as shown in Figure 3.

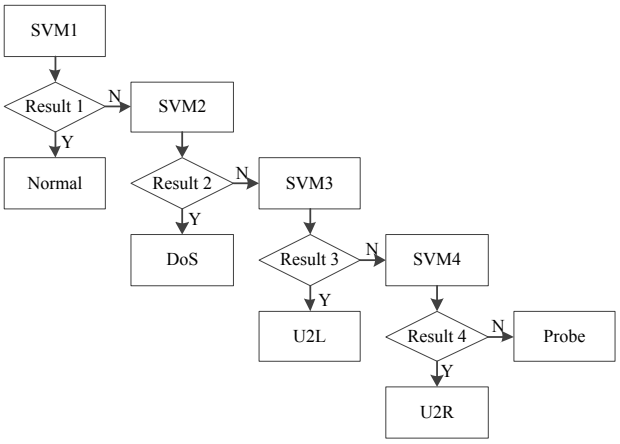

Figure 3. Classifier structure of network intrusion detection

V. EXPERIMENTAL RESULTS AND ANALYSIS

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

As the comparison model, the following indicators were used as evaluation criteria for the experimental results:

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

The simulation results were shown in Figure 4. From Figure 4, we can see that in all models, the ACO - SVM had the highest network intrusion detection accuracy, followed by GA-SVM, the lowest accuracy of network intrusion detection was BPNN, at the same time, it had the lowest false positive rate, which indicated that the ACO-SVM can accurately recognized network intrusion behavior and got the ideal detection result. At the same time, it can be seen from Figure 4b) that ACO-SVM network intrusion detection took a minimum time, which could meet the efficiency requirements of network intrusion detection,

and the superiority of network intrusion detection result was very obvious.

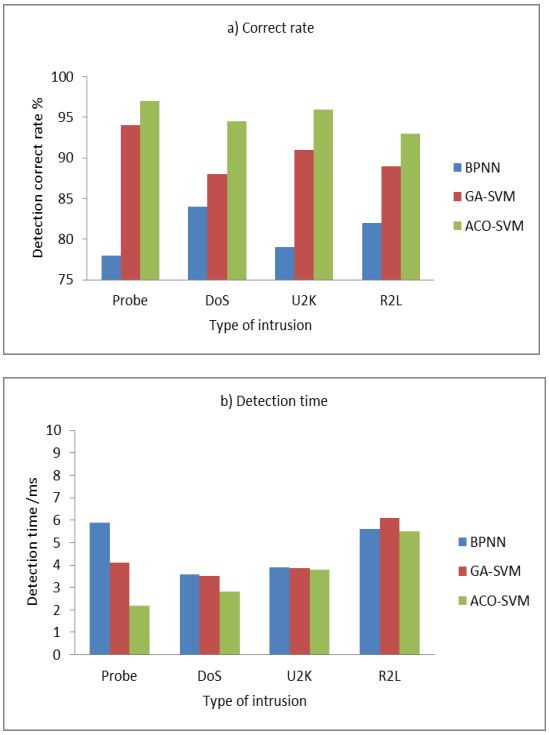

Figure 4. Comparison of network intrusion detection results

VI. Conclusion

This paper analyzed network intrusion detection in network security research, including malicious domain name modification and other network intrusion situations. In addition, SVM was used to build the model and verify the model. Compared with the traditional network intrusion model, the advantages of this model were 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.