Research on Network APT Attack Intrusion Detection Technology Based on Machine Learning Algorithm

Qingchuan Meng¹²*, YangYang³, Fengzhi Wu⁴, Xiang Chen⁵, Xiaoming Chen⁵

¹HeiLongJiang Province Cyberspace Research Center, Harbin, HeiLongJiang, 150090, China
²College of Computer Science And Technology, Harbin Engineering University, Harbin, HeiLongJiang, 150090, China
³University of Shanghai for Science and Technology, Shanghai, Shanghai, 200093, China
⁴Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, US
⁵Shenyang Aerospace University, Shenyang, Liaoning, 110136, China

*Corresponding author’s e-mail: mqchappy365@163.com

Abstract. The attack frequency of network advanced persistent threat (APT) is more and more higher, which seriously endangers the network security. In order to obtain high accuracy of network APT attack intrusion detection results, aiming at the limitations of current network APT attack intrusion detection model, a network APT attack intrusion detection model based on machine learning algorithm is proposed. A “one-to-one” network APT attack intrusion detection classifier is built through a neutral and excellent support vector mechanism of machine learning algorithm, and the current standard network APT attack intrusion detection database is adopted to verify the validity of the model. The accuracy of network APT attack intrusion detection is over 95%, and the detection error is far lower than the actual application range. It can be used in the actual network security management.

1. Introduction

With the continuous promotion of network applications, the issue of network security has attracted much attention from people. Traditional network security defense technologies mainly include data encryption and anti-virus software. These are passive defense methods, which cannot resist external intrusions, so network security cannot be effectively protected. Active network security prevention technology is mainly intrusion detection, which can detect the network security status in real time and find some illegal intrusion behaviors. Network intrusion has become the focus of current research. Among many network attacks, advanced persistent threat (APT) attack has become the most important threat to network security. It is difficult for traditional network security technologies such as firewalls to effectively protect it. Therefore, it is very important to research the APT attack intrusion prevention detection system[1-3].

In order to obtain network APT attack intrusion detection results with high accuracy, in view of the limitations of the current network APT attack intrusion detection model, a network intrusion detection model based on ant colony algorithm to determine the parameters of support vector machines was
proposed. We build a "one-to-many" network intrusion detection classifier, introduce the ant colony algorithm to determine the optimal parameters, and test the effectiveness of the model by using the current standard network APT attack intrusion detection database. The accuracy rate of network APT attack intrusion detection is over 95%, detection error is far lower than the actual application range.

2. Related theories

2.1. Support Vector Machine
Support vector machine is a kind of machine learning algorithm with excellent performance, specially for small samples, proposed by Vapnik and others. It works differently from neural network. 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 samples as far away from the optimal plane as possible. The samples on the optimal plane are called support vectors.

For a set of n samples \( \{ (x_1, y_1), \cdots, (x_n, y_n) \} \), use the function \( \varphi(x) \) to map the samples, and then classify the samples in the mapping space:

\[
f(x) = \text{sgn}(w \cdot \varphi(x) + b)
\]

Where: \( w \) is the weight, \( b \) is the threshold.

To find the optimal classification plane, we must find the optimal \( w \) and \( b \) values, and it is very difficult to obtain the optimal \( w \) and \( b \) values by directly solving equation (1). Based on the principle of structural risk minimization, the following constraints are set:

\[
y_i \cdot (w \cdot \varphi(x_i) + b) \geq 1
\]

In order to speed up the modeling speed, a relaxation variable \( \xi_i \) is used to compromise the classification accuracy and classification error, so that the optimal classification plane can be transformed into the following form:

\[
\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i
\]

The corresponding constraints is:

\[
y_i (w \cdot \varphi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \cdots, n
\]

Where \( C \) represents the degree of punishment for errors.

The Lagrange multiplier \( \alpha_i > 0 \) is introduced to obtain the dual form of equation (4):

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

And has the following constraint:

\[
\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C
\]

For the nonlinear classification problem, introducing the kernel function \( k(x_i, x_j) \), we can get:

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

Where \( k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \).

The optimal classification plane for support vector machines is:

\[
f(x) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b)
\]

Select the radial basis function, which is:
\[ k(x, x_j) = \exp(-\frac{\|x - x_j\|^2}{2\sigma^2}) \]  

(9)

Where \( \sigma \) represents the kernel width parameter.

2.2. Impact of parameters on network APT attack intrusion detection

An analysis of the working principle of support vector machines shows that the influence of parameters \( C \) and \( \sigma \) on their learning performance is very important. Select a training sample and analyze the correct rate of network intrusion detection under different parameters. The results are shown in Table 1. Analysis on Table 1 shows that even if the environment and data are the same, the intrusion detection accuracy rate of different parameters is still very different, so the optimal values of parameters \( C \) and \( \sigma \) need to be selected.

Table 1. Effect of parameters on learning performance of support vector machines

| \( C \) | \( \sigma \) | Network APT intrusion detection accuracy rate /% |
|-------|------|---------------------------------------------|
| 10    | 0.01 | 62.74                                       |
| 50    | 0.1  | 98.53                                       |
| 100   | 1    | 72.67                                       |
| 500   | 10   | 78.20                                       |
| 1000  | 100  | 95.74                                       |
| 5000  | 1000 | 67.49                                       |
| 10000 | 2000 | 77.40                                       |

2.3. Ant colony algorithm

Ant colony algorithm is a more commonly used search optimization algorithm. In the process of foraging, ants leave pheromones on the path. Other ants use pheromone to identify crawling paths. The higher the pheromone concentration, the more ants pass through the path, the higher probability that other ants choose the path[6].

If the number of ants is \( m \), the following formula can be obtained:

\[ m = \sum_{i=1}^{n} b_i(t) \]  

(10)

Where \( b_i(t) \) represents the number of ants on node \( i \).

At time \( t \), the path \((i, j)\) residual pheromone concentrations of node \( i \) and \( j \) are:

\[ \Gamma = \{\tau_{ij} | c_i, c_j \subseteq C\} \]  

(11)

Where \( \tau_{ij} \) is the pheromone concentration of \((i, j)\).

In the initial stage of ant colony algorithm, \( \tau_{ij}(0) = 0 \), the transition probability \( p_{ij}(t) \) of ant selecting the next node is:

\[ p_{ij}(t) = \begin{cases} \frac{\tau_{ij}^\alpha \eta_{ij}^\beta}{k \tau_{ij}^\alpha \eta_{ij}^\beta(t)}, & s \in allowed_k, j \in allowed_k \\ 0, & \text{otherwise} \end{cases} \]  

(12)

Where: \( \eta_{ij} \) is the local heuristic information transferred from node \( i \) to \( j \); \( allowed_k \) is the set of nodes not accessed; \( \alpha \) and \( \beta \) are the weight parameters.

After a period of time, the ant colony completes a path search and needs to update the pheromone on the path:

\[ \tau_{ij}(t + n) = (1 - \rho) \times \tau_{ij}(t) + \Delta \tau_{ij}(t) \]  

(13)
\[ \Delta \tau_{ij}(t) = \sum_{k=1}^{m} \Delta \tau_{ij}^k(t) \]  

(14)

In the formula: \( \rho \) is the volatility of pheromone; \( \Delta \tau_{ij}(t) \) is the pheromone increment on path \((i, j)\); \( \Delta \tau_{ij}^k(t) \) is the sum of pheromone, and its expression is:

\[
\tau_{ij}^k = \begin{cases} 
\frac{Q}{L_k}, & \text{Ant } k \text{ passes through } (i, j) \text{ in this cycle} \\
0, & \text{other} 
\end{cases} 
\]  

(15)

3. Construction of network intrusion model

In network APT attack intrusion detection, the optimization problem of LSSVM parameters can be expressed by the following formula:

\[
\max P(C, \sigma) \text{ s.t. } \left\{ \begin{array}{l} C \in [C_{\text{min}}, C_{\text{max}}] \\ \sigma \in [\sigma_{\text{min}}, \sigma_{\text{max}}] \end{array} \right. 
\]  

(16)

The steps for network APT attack intrusion detection are as follows:

**Step1:** Collect network status information, extract the characteristics of network APT attack intrusion detection, and process the characteristics as follows:

\[
x_i = (x - x_{\text{min}})/(x_{\text{max}} - x_{\text{min}}) 
\]  

(17)

Where \( x_{\text{max}} \) and \( x_{\text{min}} \) are the maximum and minimum, respectively.

**Step2:** Consider the support vector machine parameters \((C, \sigma)\) as a path for ant colony crawling, and model the network APT attack intrusion detection training samples according to each set of parameters to get different detection accuracy rates.

**Step3:** The ant colony’s pheromone update operation and node transfer are used to implement path crawling. Finally, the optimal parameter \((C, \sigma)\) combination is found through path optimization.

**Step4:** Establish the optimal network APT attack intrusion detection model according to the optimal parameter \((C, \sigma)\) combination.

Due to the classification of support vector machines for two categories, there are many types of network APT attacks, including mobile device attacks (MDA), malicious email attacks (MEA), and system vulnerability attacks (SVA). This article uses a "one-to-one" approach to build multiple classifiers, as shown in Figure 1.

![Figure 1. Classifier structure for network APT attack intrusion detection](image-url)
4. Experimental results and analysis
The MIT KDD Cup's network APT attack intrusion detection dataset was selected as the test object, including three types of network APT attack intrusion behaviors: MDA, MEA, and SVA. Because the size of the data set is very large, 10% of the data is randomly selected for specific experiments. In order to make the experimental results of the proposed model (ACO-SVM) convincing, a BP neural network (BPNN) genetic algorithm optimized SVM (GA-SVM) network intrusion detection model is used as a comparative model, and the following indicators are used as the evaluation criteria of the experimental results:

\[
\text{Accuracy rate} = \frac{\text{Correct number of samples detected}}{\text{Total number of samples}} \times 100\% 
\]

(a) Detection accuracy
The simulation results are shown in Figure 2. As can be seen from Figure 2 (a), ACO-SVM has the highest accuracy rate of network APT attack intrusion detection of all models, followed by GA-SVM, the lowest rate of network intrusion detection is BPNN, and the false alarm rate is the lowest. It shows that ACO-SVM can accurately recognize the intrusion behavior of APT attacks on the network and obtain ideal detection results. At the same time, it can be seen from Figure 2 (b) that the ACO-SVM network APT attack intrusion detection takes the least time and can meet the efficiency requirements of network APT attack intrusion detection. The superiority of the network APT attack intrusion detection result is obvious.

5. conclusions
The modeling of network APT attack intrusion detection is an important network security defense technology. The current intrusion detection model of network APT attack can not accurately describe the intrusion behavior of APT attack, resulting in the unsatisfactory intrusion detection results of network apt attack. Therefore, this paper proposes a machine learning algorithm based on network APT attack intrusion detection model. The support vector machine is used to fit the mapping relationship between network APT attack intrusion detection characteristics and network APT attack intrusion behavior, and a network APT attack intrusion detection model is established based on the mapping relationship. The experimental results show that the model can not only accurately identify the intrusion behavior of network APT attacks, but also the detection speed is quite fast. It obtains better intrusion detection results of network APT attacks than other models, and has a wide application prospect.

References
[1] Tang, J.C., Ming, X., Fu, S.J. (2018) A scheduling optimization technique based on reuse in spark to defend against apt attack. Tsinghua Science and Technology., 23: 550-560.
[2] Ma, Z., Li, Q. (2019) Discovering Suspicious APT Families Through a Large-Scale Domain Graph in Information-Centric IoT. IEEE Access., 7: 13917-13926.
[3] Alshamrani, A., Myneni, S., Chowdhary, A., et al. (2019) A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities. IEEE Communications Surveys & Tutorials, 21: 1851-1877.

[4] Yu, S., Li, X.F., Zhang, X.L., et al. (2019) The OCS-SVM: An Objective-Cost-Sensitive SVM With Sample-Based Misclassification Cost Invariance. IEEE Access., 7: 118931-118942.

[5] Xu, X.H., Zuo, W.M., Lin, L., et al. (2018) F-SVM: Combination of Feature Transformation and SVM Learning via Convex Relaxation. IEEE Transactions on Neural Networks and Learning Systems., 29: 5185-5199.

[6] JayaBrindha, G., Gopi Subbu, E.S. (2018) Ant Colony Technique for Optimizing the Order of Cascaded SVM Classifier for Sunflower Seed Classification. IEEE Transactions on Emerging Topics in Computational Intelligence., 2: 78-88.