Network Security Risk Prediction Based on Time-Varying Markov Model

Chao Zhou, Yajuan Guo, Wei Huang, Jing Guo & Daohua Zhu
State Grid Jiangsu Electric Power Research Institute, Nanjing 210000, China

ABSTRACT: With the application of network technology, the risk of network security is gradually increasing. In order to predict the likelihood of network risks in real-time, a Time-Varying Markov Model (TVMM) for real-time risk probability prediction was proposed. The real-time risk probability prediction method is able to predict the probability of network risk in future exactly with a real-time-updating-state probability transition matrix of TVMM. The model is used to calculate the risk probability of the network at different risk levels in network attack environment. The result shows that TVMM has higher real-time objectivity and accuracy than the traditional Markov model.
KEYWORDS: Safety risk prediction; Time-Varying Markov Model (TVMM); Network attack

1 INTRODUCTION

With the application of network technology, the characteristics of diversity, openness and connectivity brings a great impact on the network, making the network vulnerable to all kinds of threats (Wang Y. 2014). If we can predict the security risk in the network in real-time, it is very helpful to improve the security of the network (Dai Z M et al, 2012).

In order to effectively predict the network security risk, in recent years, many researchers have studied the risk prediction. The paper (Wang Y F et al, 2005) puts forward a kind of network security risk model based on immunization. The real-time risk assessment of the network system is carried out by detecting the concentration of the virtual antibody in the network system. Whereas, the method only detects the amount of the risk, but fail to predict the probability of future risk. In the paper (Liu F et al, 2008), BP neural network and average vector similarity coefficient are used to predict the risk probability of information interaction. However, due to the lack of theoretical guidance for the selection of the neural network structure, and the slow speed of learning the large sample data, the model lack real-time prediction model. Based on the grey theory, the paper (Xiaoyang L et al, 2010) puts forward a method of risk assessment under the P2P network frame. But the method has obvious system error, which results in that the prediction accuracy is not high.

Based on Markov model, this paper abandons the assumption that the state transition matrix of the system is not changed with time, and proposes a new real-time risk probability prediction model for network security attacks.

2 THE TIME-VARYING MARKOV MODEL

The traditional Markov prediction model is based on the assumption that state transition probability matrix is not changed with time (Yin Q B, 2005). However, in many practical problems, especially in the network attack environment, the state transition probability is constantly changing. So this paper updates the state transition probability matrix in real time according to the system state. Since the state transition probability is constantly changing, it can be concluded that:

$$\begin{cases} \pi(1)=\pi(0)P \\ \pi(2)=\pi(1)P' \\ \pi(3)=\pi(2)P' \\ \quad\quad ... \\ \pi(k)=\pi(k-1)P' \end{cases} . \tag{1}$$

In the formula, P' is only a symbol, which indicates the update of the state transition probability matrix from time t + 1 to t.

If an event is known at the initial state in time 0, that means that $\pi(0)$ is known. By using the recurrence formula (1), we can obtain the probability after k state transition which is $\pi(k)$. Thus we can get the probability prediction of the event at time k. There-

fore, the way to determine the state transition probability matrix $P^{'}$ is the key of prediction.

3 REAL TIME RISK PROBABILITY PREDICTION BASED ON THE TIME-VARYING MARKOV MODEL (TVMM)

Network attack is generally composed of three parts, which are information collection, attacking, and finish attacking. According to the different stages of the attack, the network risk status is divided into these: no risk state L_0 (That means the network is in a normal state), slight risk status L_1 (The network is in the state of being detected), more serious risk status L_2 (The network is under attack.) and serious risk status L_3 (The network has compromised). These states form the state space in TVMM which is: $S = \{L_0, L_1, L_2, L_3\}$. The risk status of the network is transferred as shown in Fig. 1.

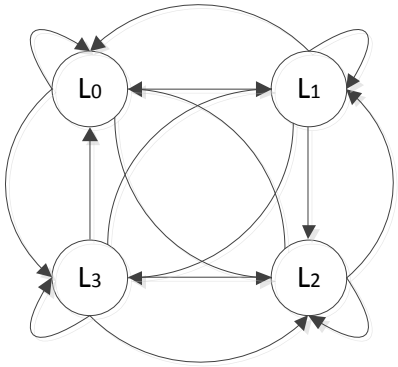

Fig. 1. State transition of the network risk

It can be determined that the transition matrix of the network risk state is:

$$P_1 = \begin{pmatrix} P_{L_0L_0}, P_{L_0L_1}, P_{L_0L_2}, P_{L_0L_3} \\ P_{L_1L_0}, P_{L_1L_1}, P_{L_1L_2}, P_{L_1L_3} \\ P_{L_2L_0}, P_{L_2L_1}, P_{L_2L_2}, P_{L_2L_3} \\ P_{L_3L_0}, P_{L_3L_1}, P_{L_3L_2}, P_{L_3L_3} \end{pmatrix}. \quad (2)$$

Calculating the state transition probability matrix, P_1 is to calculate the state transition probability from each state to any other state. The calculation of the state transition probability is generally based on frequency approximation probability.

$$p_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}. \quad (3)$$

The n_{ij} is the sample number from the state i to the state j. The way to update state transition probability matrix $P^{'}$ is shown as follow:

Step 1: According to the previous sample data and formula (2), (3), we can get the initial state transition probability matrix;

Step 2: Initialize sample data. Input data object set X and the specified number of clusters N, and in the X we select randomly N objects as the initial cluster center. Then Set the iteration termination conditions, such as the maximum number of cycles or the convergence error limit of cluster centers;

Step 3: Iterative processing. According to the similarity criterion, the data object is assigned to the nearest cluster center;

Step 4: Step 3 is repeated until the termination condition is met;

Step 5: When the new sample arrives, the cohesion degree is calculated, and we know that the sample belongs to which cluster. Combined with the last sample cluster, count the number of state transition. And then go back to the first step to re-calculate the state transition probability matrix to update.

The calculation of cohesion degree is as follows:

Step 1: The Euclidean distance is selected as the cluster partition criterion. The new sample is seen as a 3 dimensional vector, the distance of any two n dimensional vector i, and j can be calculated by using Euclidean distance. Let $d(i, j)$ represents the distance between any two vectors in the cluster G:

$$d(i, j) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (4)$$

Calculate Average distance of cluster:

$$AVG\_dis = \frac{2}{m(M+1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(i, j) \quad (5)$$

and maximum distance:

$$MAX\_dis = \max(d(d(i, j)) \quad (6)$$

So we can get cohesion degree:

$$iner(G) = \frac{MAX\_dis}{AVG\_dis} \quad (7)$$

The larger the cohesive degree is, the more similar the elements are in the cluster. According to the calculation method of state transition probability matrix, a new state transition probability matrix is obtained. Under the condition that the initial state probability is known, formula (1) is used to obtain the security risk state of the network at a certain time in the future.

4 SIMULATION TEST

In order to verify the validity of the TVMM, in this paper, CUP KDD 1999 data set (KDD99 1999) is used to do simulation test. Based on paper (DARPA 2008), Table 1 shows the corresponding risk level of various attacks according to the different stages of the attack.

Table 1. Attack classification

L0	L1	L2	L3
normal	back, land, neptune, pod, smurf, teradrop, ipsweep, nmap, satan	guess_passwd, phf, buffer_ overflow, imap, loadmodule, multihop, perl, portsweep	ftp_write, rootkit, spy, warezclient, warezmaster

Because the KDD data set is too large, we select an attack in each attack phase of the KDD for experiment. The attack combinations selected in this paper are normal, satan, guess_passwd and rootkit. The 38 kinds of attacks are classified according to Table 1. We use 2% of the normal data, all of the statan data, all of the guess_passwd data and all of the rootkit as experimental training data.

The training data are divided into four clusters C_1, C_2, C_3, C_4 , respectively representing four kinds of security risk states L_0, L_1, L_2, L_3.

Table 2 shows the results by TVMM, and Table 3 shows the traditional Markov prediction results.
From the experimental results, we can see that those who have high probability risk will have faster state transition based on the TVMM. We can also find that the test data are the same type of data at each moment form input test data. So the risk state of the network at every moment should be L_0, L_1, L_2, L_3. Because the matrix of TVMM is updated in real time with the addition of samples, the probability of the network risk will change as time goes by, which makes the prediction of risk more accurate and objective. Simulation experiments show that: At time T_1, risk probability of the network is highest in state L_0; At time T_2, risk probability of the network is highest in state L_1; At time T_3, risk probability of the network is highest in state L_2; At time T_4, risk probability of the network is highest in state L_3. The predicted results are in agreement with the actual test samples. But the traditional Markov lacks real time, which leads to the low accuracy of the results.

Table 2. The results of time-varying Markov model prediction

		L0	L1	L2	L3
{1,0,0,0}	0.99408,0.00547,0.00045,0.00000 0.00566,0.99245,0.00189,0.00000 0.05661,0.00000,0.92452,0.01887 0.10000,0.00000,0.00000,0.90000	0.99408	0.00547	0.00045	0
{0.994080,0.005470,0.000450,0}	0.99900,0.00100,0.00000,0.00000 0.25000,0.75000,0.00000,0.00000 0.00000,0.00000,0.00000,0.00000 0.00000,0.00000,0.00000,0.00000	0.99445	0.0051	0	0
{0.994453,0.005097,0,0}	0.00000,1.00000,0.00000,0.00000 0.00000,1.00000,0.00000,0.00000 0.00000,0.00000,0.00000,0.00000 0.00000,0.00000,0.00000,0.00000	0	0.99955	0	0
{0,0.999550,0,0}	0.00000,0.00000,0.00000,0.00000 0.00000,0.00000,1.00000,0.00000 0.00000,0.02127,0.95746,0.02127 0.00000,0.00000,0.00000,0.00000	0	0	0.99955	0
{0,0,0.999550,0}	0.00000,0.00000,0.00000,0.00000 0.00000,0.00000,0.00000,0.00000 0.00000,0.00000,0.00000,1.00000 0.00000,0.00000,0.12500,0.87500	0	0	0	0.9996

5 CONCLUSION

This paper presents a time-varying Markov model l for network risk prediction in real time. Compared with the traditional Markov model, this model can make up for the fact that the state transition probability matrix does not change with time. By updating the state transition probability matrix in real time, the result is more realistic, objective and accurate. The experimental results show that the proposed model is feasible and effective.

Table 3. The results of traditional Markov prediction

		L0	L1	L2	L3
{1.000000,0.000000,0.000000,0.000000}	0.99408,0.00547,0.00045,0.00000 0.00566,0.99245,0.00189,0.00000 0.05661,0.00000,0.92452,0.01887 0.10000,0.00000,0.00000,0.00000	0.99408	0.00547	0.00045	0.000000
{0.994080,0.005470,0.000450,0.000000}		0.988251	0.010886	0.000874	0.000009
{0.988251.0.010866.0.000873.0.000008}		0.982512	0.016190	0.001274	0.000024
{0.982512,0.016190,0.001273,0.000024}		0.976862	0.021442	0.001650	0.000046
{0.976861,0.021442,0.001649,0.000045}		0.971297	0.026625	0.002005	0.000073

REFERENCES

Dai Z M, Lu J J, Shan X, et al. Integrated Application of PMU Data in Control Center [J]. Jiangsu Electrical Engineering, 2012, 31(2):8-10.

DARPA. Training data attack description [EB /OL]. Http://www.ll.mit.edu /mission/communications /ist /corpora /ideval /docs /attacks. Html, 2008 - 05 - 14.

KDD99. KDD99 cup dataset [DB /OL]. http://kdd. ics. uci.edu /databases /kddcup99, 1999.

Liu F, Cai Z P, Xiao N, et al. Risk probability estimating model based on neural networks [J]. Computer Science, 2008, 35(12): 28 - 33. (in Chinese)

Wang Y. A Design of Gaming Theory Based Defense System for Power System Cyber-Security [J]. Jiangsu Electrical Engineering, 2014, 33(5):82-84.

Wang Y F, Tao L I, Xiao-Qin H U, et al. A Real-Time Method of Risk Evaluation Based on Artificial Immune System for Network Security [J]. Acta Electronica Sinica, 2005, 33(5):945-949.

Xiaoyang L, Cai F, Jiandong Y, et al. Grey Model-Enhanced Risk Assessment and Prediction for P2P Nodes[C]// International Conference on Frontier of Computer Science and Technology. 2010:681-685.

Yin Q B. Research on Technology of Intrusion Detection Based on Linear Prediction and Markov Model [J]. Chinese Journal of Computers, 2005.