Analysis of Network Security Situation Based on Principal Component Analysis and Phase Space Reconstruction

Wenzhi ZHU[1]

[1]Information Management Department, Zhongnan University of Economics and Law, Wuhan, 430073, China

Keywords: Network Potential Hazard Trend Estimation; Estimation; Grey Relational Analysis; Support Vector Machine

Abstract. In order to improve the accuracy of network potential hazard trend estimation, a method of network potential hazard estimation based on the combination of grey relational analysis (GRA) and improved support vector machine (SVM) is proposed. At first, determine evaluation index weight by GRA, then optimize SVM parameter by particle swarm optimization (PSO) to establish network potential hazard trend estimation model, and finally, test the model's effectiveness by simulation experiment.

Introduction

At present, network intrusion emerges one after another, and the network information security faces huge threat. The problem of network security attracts more and more attention [1]. Network security detection system is a passive defense, difficult to accurately grasp the security status of the network system, but network potential hazard trend estimation can monitor the network security in real time, and have a quick and precise judgment on security situation. It is a main focus in today's network security research, because it makes up the traditional prevention methods [2].

There are mainly two methods in network potential hazard trend estimation, qualitative and quantitative estimation method. The qualitative method includes comparative method, Delphi, fuzzy theory and etc [3,4], whose estimation results are not objective for its strong subjectivity. The quantitative method includes factor analysis, cluster analysis, SVM, neural network and etc [5-7], of which SVM is today's main network potential hazard trend estimation model for its advantages including small sample and good generalization ability [8]. But there are two problems needing to be solved in the application of SVM. The first is to determine the index weight of network potential hazard trend estimation. The second is to optimize the SVM parameter [9]. GRA is a multivariate statistical analysis method which can objectively determine index weight, while the particle swarm optimization (PSO) has fewer parameters, and the capability of global searching. Therefore, introducing them to network potential hazard trend estimation can enhance the credibility of situation assessment results [10].

Selection and Quantification of the Index Set of Network Potential Hazard

Network security is a dynamic system which is affected by many factors. Its uncertainty and randomness lead to the complexity of network security evaluation. Network security is not only affected by the attack frequency, the number of attack source, but affected by the occupancy rate of bandwidth and other factors. The influence of different factors on the network security is different in degree and time, which causes a complicated non-linear relationship between network security and its influence factors. As a result, the network security evaluation index should be chosen accurately.

Index selection of the network potential hazard trend estimation

In the establishment of the network potential hazard trend estimation index system, complex dynamic characteristics of network should be taken into consideration. According to the network potential hazard trend estimation principle, 6 estimation indexes closely related to network security

are selected to form the evaluation index set of the model, including the priority of attack type, number of attack source, attack frequency, degree of importance in time, the occupancy rate of bandwidth and the degree of importance in host, namely: U={the priority of attack type, number of attack source, attack frequency, degree of importance in time, occupancy rate of bandwidth and degree of importance in host }

Index quantification of the network potential hazard trend estimation

(1) In a certain period, the more frequent attacks are, the greater the threat is, and in order to reduce the impact of attack frequency to threat to [0.0,1.0), the specific quantification is as follows:

$$y(f) = f/(1+f),\ f \in [0,\infty),\ y \in [0,1) \tag{1}$$

In the formula, f is defined as the number of attacks of the same type in unit time

(2) Usually for a coordinated attack, network is threaten more greatly as the number of attack source becomes larger, and in order to reduce the impact of the number of attack source to threat to the range of [0.0, 1), the quantitative formula is as follows

$$y(u) = u/(1+u), u \in [0,\infty),\ y \in [0,1) \tag{2}$$

(3) Considering risk level characteristics of network security estimation, V, the evaluation set of the model, is classified as 5 levels, Very Low, Low, Moderate, High, Very High, namely,

$$V = \{VL, L, M, H, VH\} \tag{3}$$

Finally, the network security evaluation criteria are shown in table 1.

Table 1 Quantitative criteria network situation security evaluation index

Evaluation index	VL	L	M	H	VH
Threat level of attack	3	2.5	2.0	1.5	1.0
Number of attack source	0	1	2	3	4
Attack frequency	1	5	10	15	20
Degree of importance in time	1	2	3	0	0
occupancy rate of bandwidth	0	1	2	3	4
degree of importance in host	0	2	4	6	8

Establishment of matrix for the network potential hazard trend estimation

Firstly, each index ui is graded according to vj in the evaluation set V, and then rij, the value of vj corresponding to ui, the jth element in V, is calculated according to the following formula

$$r_{ij} = n_{ij} / n_{总} \tag{4}$$

In the formula, nij is the expert number which makes the evaluation to ui, the ith index in index set, and nsum is the sum of experts.

Thus the evaluation vector of index ui, can be worked out.

$$r_i = (r_{i1}, r_{i2}, \cdots, r_{ik}) \tag{5}$$

Through the above method, we can know that n indexes have n evaluation vectors, $r_1, r_2, \cdots r_n$. According to this, a mapping relation can be determined

$$R = (r_1, r_2, \cdots r_n) \tag{6}$$

1 The matrix for the network potential hazard trend estimation is

$$R = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nk} \end{pmatrix} \tag{7}$$

Grey Relational Analysis Determines the Weight of Index Set

Set Xi as the system evaluation index, and its observation data as xi(k), k=1,2,..., n, then $X_i = \left(x_i(1), x_i(2), \cdots, x_i(n)\right)$ is called the behavior sequence of the evaluation index *Xi*, and the

method of grey relational analysis can be divided into the following steps:

(l) Set m data sequences to form the following matrix:

$$\left(X_1^{'}, X_2^{'}, \cdots, X_n^{'}\right) = \begin{pmatrix} x_1^{'}(1) & \cdots & x_n^{'}(1) \\ \vdots & \ddots & \vdots \\ x_1^{'}(m) & \cdots & x_n^{'}(m) \end{pmatrix} \tag{8}$$

In the formula, n is the number of index, $X_i^{'} = \left(x_i^{'}(1), x_i^{'}(2), \cdots, x_i^{'}(m)\right)^T$, $i = 1, 2, \cdots, n$.

(2) Reference data are listed as

$$X_0^{'} = \left(x_0^{'}(1), x_0^{'}(2), \cdots, x_0^{'}(m)\right) \tag{9}$$

(3) Nondimensionalize index data as following

$$\left(X_0, X_1, \cdots, X_n\right) = \begin{pmatrix} x_0(1) & \cdots & x_n(1) \\ \vdots & \ddots & \vdots \\ x_0(m) & \cdots & x_n(m) \end{pmatrix} \tag{10}$$

$$x_i(k) = \frac{x_i^{'}(k)}{\frac{1}{m}\sum_{k=1}^{m} x_i^{'}(k)} \tag{11}$$

$$x_i(k) = \frac{x_i^{'}(k)}{x_i^{'}(1)} \quad i = 0, 1, \cdots, n; k = 1, 2 \cdots, m \tag{12}$$

(4) Compare the absolute difference between the corresponding elements of each sequence and reference sequence, namely

$$\left|x_0(k) - x_i(k)\right| \quad (i = 1, \cdots, n; k = 1, 2 \cdots, m) \tag{13}$$

(5) Determine $\min_{i=1}^{n} \min_{k=1}^{m} \left|x_0(k) - x_i(k)\right|$ and $\max_{i=1}^{n} \max_{k=1}^{m} \left|x_0(k) - x_i(k)\right|$.

(6) Calculate respectively the correlation coefficient of corresponding elements of each comparative sequence and reference sequence, namely

$$\xi_i(k) = \frac{\min_i \min_k \left|x_0(k) - x_i(k)\right| + \rho \bullet \max_i \max_k \left|x_0(k) - x_i(k)\right|}{\left|x_0(k) - x_i(k)\right| + \rho \bullet \max_i \max_k \left|x_0(k) - x_i(k)\right|} \tag{14}$$

In the formula, $k = 1, 2 \cdots, m$, ρ is a resolution ratio, taking its value in (0,l). The smaller ρ is, the higher the distinguishing ability of correlation coefficient is.

(7) Calculate average correlation coefficient. Calculate respectively mean value of correlation coefficient of corresponding elements in *n* indexes and reference sequences, so as to reflect the correlation of reference sequence and comparative sequence, which can be called average correlation coefficient, set as r_{0i}, then

$$r_{0i} = \frac{1}{m}\sum_{k=1}^{m} \xi_i(k) \quad i = 1, \cdots, n \tag{15}$$

(8) Based on average correlation coefficient, evaluation results can be worked out.

(9) Based on the results from grey relational analysis, the weight vector of each index can be worked out by normalizing vector $r = (r_{01}, r_{02}, \cdots, r_{0n})$, composed by r_{0i}, as following

$$W = (w_1, w_2, \cdots, w_n) \tag{16}$$

Network Potential Hazard Trend Estimation Model

Support vector machine

Given dataset: (x_i, y_i). According to risk minimization principle, the SVM optimal hyperplane

is represented as:

$$y = w^T \varphi(x) + b \tag{17}$$

In the formula, w is normal vector of hyperplane, while b is offset vector of hyperplane.

If it is a nonlinear classification problem, then it can be transformed into quadratic optimization, namely,

$$\min J(w,\xi) = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}\xi_i \tag{18}$$

the corresponding constraint condition is

$$\begin{gathered} y_i(w \bullet \Phi(x_i) + b) \geq 1 - \xi_i \\ \xi \geq 0, i = 1, 2 \cdots, n \end{gathered} \tag{19}$$

In the formula, $\xi = (\xi_i, ..., \xi_l)^T$, c is penalty parameter.

For the classification of large sample, SVM learns slowly. By introducing Lagrange multiplier, SVM classification problem can be transformed into dual problem, which can solve hyperplane optimization problem and accelerate classification speed, thus working out SVM decision function:

$$f(x) = sign(\sum_{i=1}^{n}\alpha_i y_i(\varphi(x) \bullet \varphi(x_i)) + b) \tag{20}$$

In the formula, $sign$ is sign function, while α_i is Lagrange multiplier.

Use different kernel function to create different support vector classifiers. Radial basis function (RBF), polynomial kernel function, Sigmoid function and other kernel function are commonly used. Because RBF is helpful to parameter optimization for it should only ascertain one parameter (namely, width parameter of kernel function, σ), it chooses RBF to construct SVM. The definition of RBF kernel function is as follows:

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|}{2\sigma^2}\right) \tag{21}$$

SVM parameter influences its learning ability and generalization ability. So the key point for RBF kernel function SVM to choose optimal SVM parameter includes adjustment parameter C and kernel width σ.

Particle swarm optimization (PSO)

In Particle Swarm Optimization (PSO), particles have initial position and speed. Its quality is determined by fitness function. The optimal solution of the problem corresponds to the "food" in the search space. In the flying course of particle swarm, each particle decides its next step in the direction and distance to travel according to the current optimal particle and its own memory, thus finding “food”, namely the optimal solution. The position and speed of particle i can be respectively indicated as *Xi(t)* and *Vi(t)*. In each iterative process, the particle can upgrading its own speed and position by tracing the optimal solution found by itself and the whole particles, namely personal best value (*Pbest*) and global best value (*gbest*), which can be specified as

$$V_{ik}(t+1) = wV_{ik}(t) + c_1 r_1 (P_{ik}(t) - X_{ik}(t)) + c_2 r_2 (P_{gk}(t) - X_{ik}(t)) \tag{22}$$

$$X_{ik}(t+1) = X_{ik}(t) + V_{ik}(t+1) \tag{23}$$

In the formula, w is inertia weight; c1 and c2 are acceleration constants. In general, c1=c2=2, and the value of r1 and r2 is random number ranged from 0 to 1.

Steps of PSO-SVM parameter

(1) Based on problem scale, particle swarm scale, *m*, is determined and particles are initialized, and two-dimensional vector parameter of particles is represented by C and σ, to determine the largest number of iterations and other parameter.

(2) SVM is used to model training samples, and to calculate fitness value of particles.

(3) Compare and upgrade the best value of the individual particle and the whole particles.

(4) Upgrade every particle according to formula (22) and (23)

(5) Make a judgment on termination criteria. If termination criteria are met, then return to step

(6), otherwise, add iterations and return to step (2) to continue optimization.

(6) Inverse-coding optimum particle position to get optimum value of SVM parameter C and σ.

Simulation Experiment

Data sources

To guarantee the performance of the network potential hazard trend estimation model of this study, algorithm is achieved through programming tool, matlab 2007 in the hardware environment of P4, 3.0G, CPU, 1G memory and the operating system known as Windows XP. Experimental data come from DARPA99 intrusion detection data set of MIT Lincoln Laboratory [11].

Model performance evaluation index

Attack energy is adopted in model performance evaluation index. According to network threat situation evaluation principle, the function calculating network attack energy is

$$E = \sum F(H+1)T / A \tag{24}$$

In the formula, attack frequency, degree of importance in time, degree of importance in host, and attack threat level are respectively represented as F, T, H and A.

Result analysis

Ascertain the weight of each evaluation index through GRA, and estimate the network potential hazard trend through PSO–SVM. Network potential hazard trend estimation result is represented in Fig. 1. And according to the energy value, the changing curve for network attack energy is made as Fig.2. Comparing Fig. 1 and Fig. 2, the two curves of network security threat situation value and network attack value, achieved by modeling, are basically identical, which can prove the reasonability, effectiveness and accuracy of the results achieved by network security threat situation evaluation algorithm presented in the study.

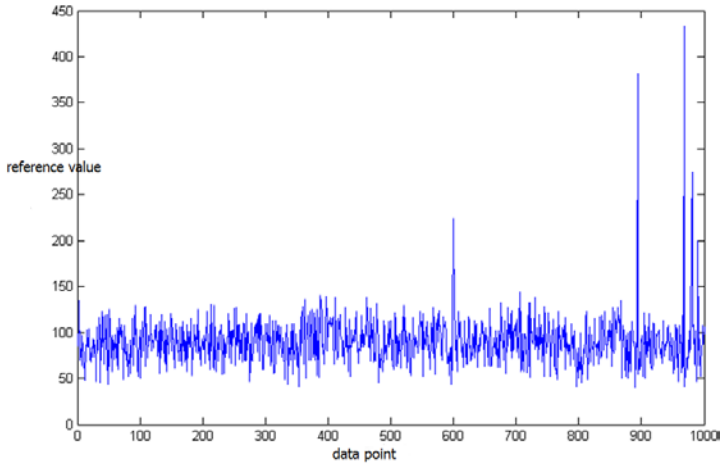

Fig.1. tendency chart of network potential hazard

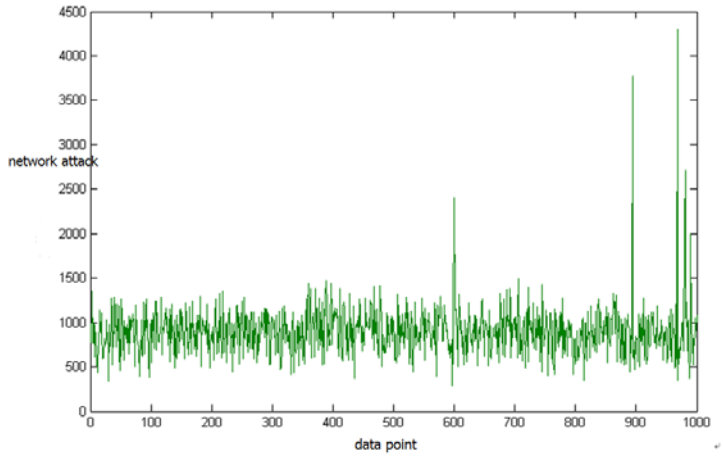

Fig.2. Changing curve of network attack energy

Conclusion

Network potential hazard trend estimation can monitor network security changing trend, finding its potential and possible network attacks and then taking effective precautionary measures to improve network security. Aiming at the existing problems in the current network potential trend estimatiion, the paper, taking advantage of grey relational analysis (GRA) and support vector machine (SVM), presents a new network potential trend estimation model based on the combination of GRA and SVA. The simulation result indicates that the model can seize the overall changing trend of network security and have an acurate and objective evaluation on the network potential trend estimation. The evaluation results are helpful to guide network administrators to take corresponding measures for future network security incidents.

References

[1] Hong-Tao X U, Wang Y G. Noise cancellation for telemetry signal based on reconstructed phase space and principal component analysis: Noise cancellation for telemetry signal based on reconstructed phase space and principal component analysis[J]. Journal of Computer Applications, 2010, 30(3):793-795.

[2] Gamo C, Gaydecki P, Zaidi A, et al. Principal Component Analysis of Network Security Data Based on Projection Pursuit[M]// Network Computing and Information Security. 2012:380-387.

[3] Tian Z D, Shu-Jiang L I, Wang Y H, et al. Network traffic prediction method based on KPCA optimized ESN[J]. Electric Machines & Control, 2015.

[4] Luan J, Bi G, Wang H, et al. Principal component analysis and identification of power quality disturbance signal phase space reconstructed images[C]// Control Conference. IEEE, 2012:5229-5233.

[5] Shi Y, Peng X, Zhang W, et al. A chaotic characteristics identification method for network security situation time series[J]. Journal of Information & Computational Science, 2012, 9(5):1309-1319.