Information System Security Situation Assessment Based on Data Mining

Qianjin Zhang

School of Information Technology, Anhui Vocational College of Defense Technology, Lu'an, Anhui 237011, China

Abstract

One of the most critical issues in security issues have been information system security situation estimation accuracy estimation of the traditional information system can not meet the actual application of the information system, in order to estimate the effect of improving security situation information system, put forward the estimation model of information system based on data mining security situation. The method of collecting a lot of information system security of the historical data, and then using data mining combined kernel function to dig out the characteristics of information system security situation from the historical data, the establishment of information system security situation assessment model, finally model were compared, and the other information system security situation assessment results show that the model can be found to change the trend of information system security situation, get information system security situation better contrast model estimation results, the estimation results can improve the security of the information system and formulate corresponding preventive measures.

Keywords

Information system; data mining; security situation; estimation model.

1. Introduction

With the development of information technology, every industry has its own information system, and information security is a key technology of information system. Due to different security awareness of managers and diversified illegal attack means, it is increasingly important to accurately predict the security situation of information system [1-3].

The current information system security situation prediction mainly USES the time series forecasting method is studied, first to collect the information security data of different periods, and then USES the data mining techniques such as autoregressive moving average (ARIMA), neural network, the relevance vector machine (RVM), and other [4-7] for information system security situation prediction model is established, in all models, ARIMA belongs to the linear model, and the information system security situation belongs to nonlinear prediction problem, thus ARIMA to smooth processing of data, often lead to information system security situation prediction accuracy is low [8, 9]. For ARIMA, the nonlinear prediction ability of neural network is stronger, but it requires a large number of training samples and slow convergence speed, leading to the unreliable prediction results of information system security situation [10-12]. The predictive performance of RVM is better than that of neural network, but its performance is directly related to the types and parameters of kernel functions. A single kernel function is used to establish the security situation prediction model of information system, and the prediction accuracy needs to be further improved [13].

In order to improve the predictive accuracy of information system security situation for the deficiency existing in the current information system security situation prediction model, and combining with the characteristics of the change of the information system security situation and put forward the model of information system security situation prediction based on data mining (ARIMA - RVM), using ARIMA and the advantage of RVM, information system security situation prediction model is set up, respectively, of information system security situation forecast trend item and random item modeling and prediction, the prediction results are superimposed on them get the final prediction

results, the information system security situation and to test its performance by simulation experiments.

2. Data Mining Technology

2.1 Wavelet Transform

The security situation of information system has the characteristics of periodic and random changes, which cannot be accurately described by a single model. Therefore, wavelet transform is adopted to decompose the security situation, and trend term and random term are obtained, and then respectively modeling and forecasting are carried out. The specific working steps are:

The j layer decomposition of the original information system security situation data (W) was carried out by using wavelet transform, and the trend term (Xj+1) and random term (XZj+1) were obtained as follows:

$$\begin{cases} X_{j+1} = HX_j \\ Z_{j+1} = GZ_j,\ j = 0,1,...,J \end{cases} \tag{1}$$

Where, H and G are respectively high-pass and low-pass filters.

Finally, the prediction results of the information system security situation of ARIMA and RVM are given.

$$x_j = H^* x_{j+1} + G^* z_{j+1},\ j = J-1, J-2,...,0 \tag{2}$$

Where, H^* and G^* represents the dual operator of sum.

2.2 ARIMA

ARIMA is a classical linear modeling method, which can model the trend term of information system security situation. The main steps are:

(1) The security situation data of the information system is non-stationary. First, the security situation data of the information system is differentiated and changed into stable data. The stationary data after the difference processing is $\{x_t^{'}\}$.

(2) Calculate the autocorrelation and partial autocorrelation functions according to the security situation sample of information system, and establish the preliminary model type of security situation prediction of information system. The ARIMA model of $\{x_t^{'}\}$ can be expressed as:

$$x_t^{'} = \sum_{j=1}^{p} a_j x_{t-j}^{'} + \sum_{k=0}^{q} b_k e_{t-k} \tag{3}$$

(3) The value of the parameter (p, q) in equation (3) is predicted using AIC.

(4) The ARIMA model of information system security situation prediction was established by using matrix prediction method to determine the value of equation (3) parameter (a_j, b_k).

2.3 RVM

Suppose the sample of security situation training of information system is: $\{x_i, t_i\}_{i-1}^{n}$, (i=1,2,3,…,n), xi, ti are the security situation input and output of information system respectively. Then RVM can describe the relationship between them.

$$t_i = y(x_i, \omega_i) + \omega_i \tag{4}$$

In the equation, axial I represents the noise contained in the data.

By introducing the kernel function K(x,xi), we can get:

$$y(x; w) = \sum_{i=1}^{n} w_i K(x, x_i) + w_0 \tag{5}$$

Where, w represents the weight vector.

The probability distribution of the i-th training sample is calculated as:

$$P(ti|xi)=N(ti|y(xi;w),\sigma^2) \tag{6}$$

The maximum likelihood function of the training sample for the security situation of all information systems is established by using the superparameterβ.

$$p(t \mid w,\beta)=(\frac{\beta}{2})^{N/2}\exp\left\{-\frac{\beta}{2}\|t-\varphi w\|^2\right\} \tag{7}$$

Where, t=[t0,t1,…,tN]T, $\varphi \in R^{N\times(N+1)}$, represents the matrix.

a_j^{-1} represents the variance of wj, so its gaussian distribution is defined as:

$$p(w \mid a)=\prod_{j=0}^{n} N(w_j \mid 0, a_j^{-1}) \tag{8}$$

The posterior distribution of w is calculated by bayesian formula:

$$p(w \mid t,a,a)=\frac{p(w \mid a)p(t \mid w,\beta)}{p(t \mid a,\beta)} \tag{9}$$

By changing equation (9), we can get:

$$p(w \mid t,a,a)=N(w \mid \mu,\sum) \tag{10}$$

Where, the calculation formulas of ∑ and μ are respectively:

$$\sum=(\beta\varphi^T\varphi+A)^{-1} \tag{11}$$

$$\mu=\beta\sum\varphi^T t \tag{12}$$

Where, A represents the diagonal matrix.

The values of aj and β are:

$$a_j=\frac{1}{\mu_j^2+\sum_j}\frac{\gamma_j}{\mu_j^2},\ j=0,\ 1,\ldots,\ n \tag{13}$$

$$\beta=\left(n-\sum_{j=0}^{n}(1-a_j\sum_{jj})\right)/\|t-\varphi\mu\|^2 \tag{14}$$

Where, μj is the JTH element ofμ, and ∑jj is the JTH diagonal element of ∑.

For the new input information system security situation sample x*, its predicted value (t*) can be obtained through equation (15).

$$t_*=\varphi(x_*)\mu \tag{15}$$

RVM kernel function directly influences the security situation prediction results of information system. Currently, there are many kernel functions that meet the conditions. The most commonly used polynomial kernel function and RBF kernel function are:

$$K_{\text{poly}}=((x\cdot x_i)+1)^d \tag{16}$$

$$K_{rbf}=\frac{\exp(-|x-x_i|^2)}{2\sigma^2} \tag{17}$$

Single kernel function has obvious defects. Therefore, this paper constructs combined kernel function to realize the security situation prediction of information system.

$$K_{\text{mixed}}=\rho_1 K_{\text{poly}}+\rho_2 K_{rbf} \tag{18}$$

Where, ρ1 and ρ2 are weights.

3. Information System Security Situation Prediction Model of Data Mining Technology

The security change of information system is very complex, because it is affected by many factors. Therefore we use wavelet transform to information system security situation data processing, get the information system security trend item and random item, then ARIMA to forecast the trend of the information system security situation in item, at the same time using SVM to predict random item, of the information system security situation in the final to overlay information system security situation prediction results, the principle is shown in figure 1.

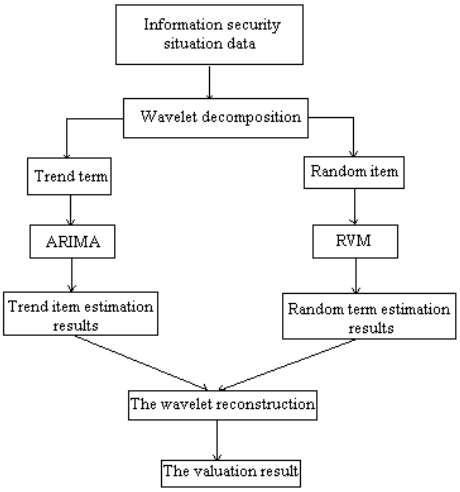

Figure 1. Working principle diagram of ARIMA-RVM

4. Results and Analysis

In order to test the information system security situation prediction model of arima-rvm, the information system security data of one enterprise was selected as the research object, the first 200 samples were selected as the training sample set, and the other samples were used to test the generalization ability of the model, as shown in figure 2. In addition, ARIMA and RBF kernel function RVM (RBF-RVM) were selected for comparison test of information system security situation prediction. Root mean square error (RMSE) was used to evaluate the prediction effect of information system security situation. The calculation formula is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \tag{19}$$

Where, x_i is the value of the security situation of the original information system, $\hat{x}_i$ is the predicted value of the security situation of the information system, and n is the sample number.

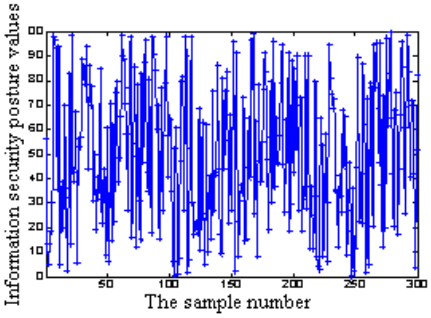

Figure 2. Sample security situation of information system

The security situation data of the first 200 information systems were input into RVM for training. Different kernel functions were used to predict other samples, and their RMSE was counted. The results were shown in table 1. Information system security situation prediction error of table 1 compare and analysis found that the polynomial kernel function of information system security situation prediction error is the largest, this is because it cannot fully describe the characteristic of information system security situation changes, and the combination of kernel function of information system security situation prediction error is smaller than a single kernel function, the results show that the selection combination kernel function of RVM to establish information system security situation prediction model is feasible.

Table 1. RMSE comparison of different kernel functions

Kernel function type	RMSE
Polynomial kernel function	4.95
RBF kernel function	3.03
Combinatorial kernel function	2.15

In order to further improve the prediction accuracy of information system security situation, arima-rvm makes predictions on it, and the results obtained are shown in figure 3. At the same time, ARIMA and rbf-svm are adopted to model and predict the security situation of information system, and their predicted results are shown in figures 4 and 5. To figure 3 ~ 5 of the information system security situation analysis, the predicted results can be found that ARIMA - RVM prediction results of the information system security situation in the optimal, can better simulate the future trend of the information system security changes, mainly because of the information system security situation trend part by ARIMA capture, overcome the limitations of single model, improve the prediction precision of the information system security situation, the single ARIMA and RBF - the prediction accuracy of the information system security situation in RVM is low, there are many points of prediction error is still large.

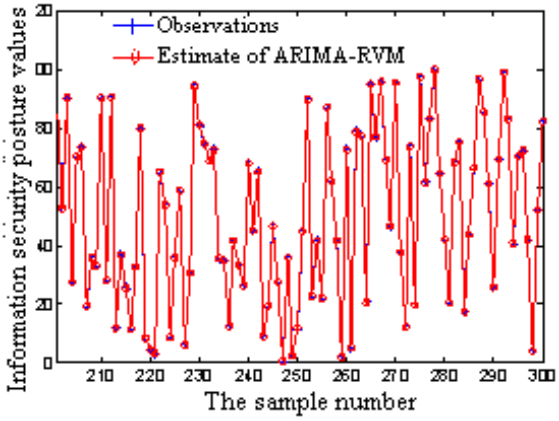

Figure 3. Prediction results of ARIMA-RVM information system security situation

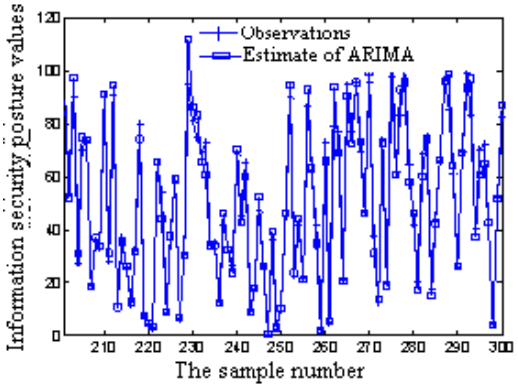

Figure 4. Prediction results of ARIMA information system security situation

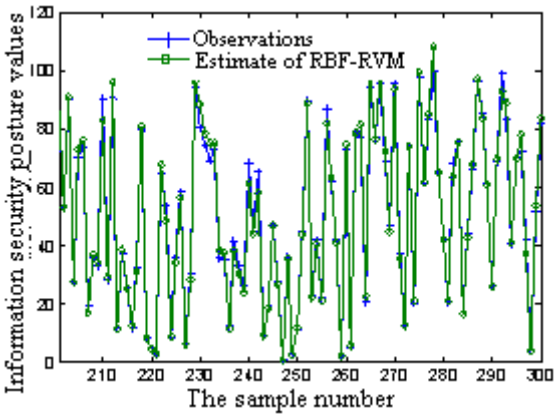

Figure 5. Prediction results of RBF-RVM information system security situation

5. Conclusion

Prediction method in traditional information system security situation prediction accuracy is low, can't meet the problem of information system application, puts forward the model of information system security situation prediction based on data mining, the results show that this model can be found that the tendency of the information system security situation, better than other models of information system security situation prediction results, has good practical application value.

References

[1] Y. Zhang, X.B. Tang, X.L. Cui, H.S. Xi: Network Security Situation Awareness Approach Based on Markov Game Model. Journal of Software, Vol.22(2011)No.3, p.495-508.(In Chinese)

[2] Y.Jia, X.W. Wang, W.H. Han, A.P. Li, W.C. Cheng. YHSSAS:Large-scale Network Oriented Security Situational Awareness System. Computer Science, Vol.38(2011) No.2, p.4-8. (In Chinese).

[3] R.Z. Xu, T.H. Chang, G.j. Lv. The Research on Prediction Method of Network Security Posture Based on Time Series. Mathematics in Practice and Theor, Vol.40(2010) No.12, p.124-133.

[4] J. Meng, C. Ma, J.L. He, H. Zhang. Network Security Situation Prediction Model Based on HHGA-RBF Neural Network.Computer Science, Vol.38(2011) No.7, p.70-73.

[5] G. Wang, J.H. Zhang, N.Wu. Application Research on Network Security Situation Prediction method. Computer Simulation, Vol.29(2012) No.2,p.98-101.

[6] Z.Q.Tang, X.W. Chen, H.T. Dai, F. Guo. Research of information security risk assessment based on multiple attribute group decision-marking theory. Computer Engineering and Applications, Vol.47(2011) No.15, p.104-106.

[7] F. Wang, M.K. Huo, X.T. Wang. Information System Security Risk Assessment Based on the Fuzzy Gray-level and Countermeasures. Information Science, Vol.32(2014) No.1, p.110-114.

[8] Z.Z. Wang, X. Jiang, X.Y. Wu, X.Tan. Planning Exploitation Graph-Bayesian networks Model for Information Security Risk Frequency Measurement. ACTA Electronica Sinica, Vol.38(2010) No.2, p.18-22.

[9] Y. Fu, X.P. Wu, Y.X. Song. Appl ication of fuzzy reasoning and multi-layer neural network to security risk assessment of information system. Journal of Naval University of Engineering, Vol.23(2011) No.1, p.: 10-16.

[10] F.W.Li, B.Zheng, J. Zhu, H.B. Zhang. A method of network security situation predictionbased on AC-RBF neural network. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), Vol.26 (2014) No.5, p.576-583.

[11] H.Ruan, D.P. Dang. Risk assessment of information security based on RBF fuzzy neural network.Computer Engineering and Design, Vol.32(2011) No.6, p.2113-2115.

[12] D.P. Dang, Z. Meng. Assessment of information security risk by support vector machine. Journal of Huazhong University of Science and Technology(Nature Science Edition), Vol.38(, 2010) No.3, p.46-49.
[13] J.G. Lou, J.H. Jiang, Z.G. Shen, Y.L. Jiang. Software Reliability Prediction Modeling With Relevance Vector Machine. Journal of Computer Research and Development, Vol.50(2013) No.7, p.1542-1550.