Network security situation prediction based on singular spectrum analysis

Jizhi Wang[1,2], Bo Hu[2], Jianguo Jiang[2], Jiqiang Liu[3]

1. Key Laboratory for Computer Network of Shandong Province, Shandong Computer Science Center (National Supercomputer Center in Jinan), China

2. Institute of Information Engineering, Chinese Academy of Sciences, China

3. Beijing Jiaotong University, China

wangjzh@sdas.org, hubo@iie.ac.cn, jiangjianguo@iie.ac.cn, jqliu@bjtu.edu.cn

Key Words: Network Security, Situation Prediction, Singular Spectrum Analysis, ARMA Model

Abstract: Network security situation prediction based on historical data is to provide warning in advance with network administrator. However, the current proposed approaches ignores the fact that the original time series of security situation include the random component which is not able to be forecast. Singular spectrum analysis is utilized to separate the random component. After that, ARMA model is used to predicate the security situation value in near future. The experiment shows the prediction benefits from the separation.

1 Introduction

With the rapid development of information technology, many services can be provided through Internet, such as sharing resources, that make the public's life and work convenient. Almost all the organizations have office networks which is connected with Internet. In the condition, network security issues have emerged, which has become a key issue of network management. On every day, a network, especially Internet Infrastructure, may be attacked by malicious hackers. A network administrator is often confused by thousands of network security events, who is more interested in viewing the overall network security situation, that is called situation awareness[1]. Situation prediction can forecast incoming network security situation based on historical situation awareness information. With the help of the prediction, a network administrator can take some measures to resist network attack in advance. Thus, network security situation prediction attracts many researchers[2].

In current research, many methods for situation prediction have been proposed. These network security situation prediction techniques can be grouped into three categories: time series analysis[3-4], neural network[5-6], and gray theory[7]. In addition, some combined methods[8] were proposed.

However, there is a basic problem that whether the security situation is predicable. In fact, the security situation contains random components which is not able to be predicable. So before the historical situation data is used to predicate, it need to be pre-processed to drop the random components because predicating random components is no sense.

Thus, a pre-processing method based on singular spectrum analysis (SSA)[9] is proposed in the paper. It can be used to separate random components from the situation data and acquire better predicated result.

2 The problem

The current proposed methods for network security prediction are based on raw data of security situation. Taking the following example, we show why this kind of method is questionable.

The raw security situation data were collected from the network of some an organization from 14:00 of the first day to 14:00 of the second day. The security situation values were computed every half an hour, totally 48 data points which show in Fig.1. These values are normalized between 0 and 1, where 0 denotes very secure situation and 1 denotes very serious situation.

© 2016. The authors - Published by Atlantis Press

From Fig.1, it is shown that the security situation was serious from 19:00 (the 10th data point) of the first day to 3:00(the 26th data point) of the second day, especially the most serious case between 22:30 (the 17th data point) and 2:30 (the 25th data point). Obviously, the network is under more attack during midnight than other of daytime.

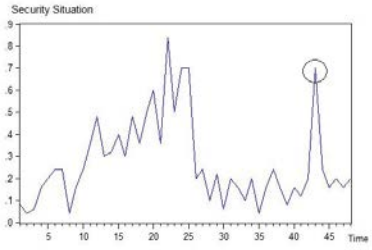

Fig.1 the Security Situation of a network during 24 hours

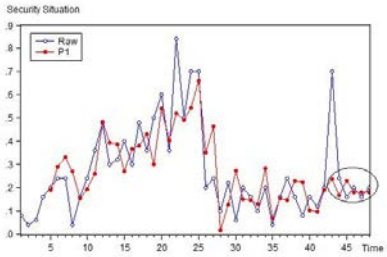

Fig.2 The prediction result based on the first 42 data point

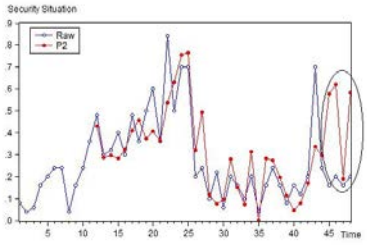

Fig.3 The prediction result based on the first 44 data point

The 43th data point (11:30, circled in Fig.1) is another peak value. Through manual inspection, the peak value is made by a short-time DoS (Denial of Service) attack from 11:00 to 11:10. From the point of the security situation prediction's view, the unexpected situation can hardly be forecast because attackers randomly decided when to attack. In other words, the kind of sudden change is hardly forecast.

To prove the issue, we use the first 42 data point to forecast next 6 data point with the help of ARMA model[5] (Auto-Regressive and Moving Average Model). The result is shown in Fig.2 and Table 1.

Table 1 Raw data and prediction results

Time	42	43	44	45	46	47	48
Raw	0.2	0.7	0.24	0.16	0.2	0.16	0.2
P1		0.234	0.167	0.23	0.179	0.179	0.178
P2				0.576	0.62	0.19	0.582
P3				0.17	0.165	0.156	0.144

In Table 1, the row of "Time" denotes the sequence number of data point, from the 42nd data point to the 48th data point. The row of "Raw" denotes the raw data of security situation. The raw of "P1" denotes the prediction results based on the first 42 data points.

In Fig.2, the prediction results are circled. Obviously, the peak value of the 43rd data point is not forecast though the prediction results of data points from the 44th to the 48th are of valuable reference. Thus, it is shown that these temporary high-effect attack events are not able to be forecast exactly.

On the other hand, what will happen if the predication is based on data including the peak value, like the 43rd data point? Then, we use the first 44 data point to forecast next 4 data point with the help of ARMA model. The result is shown in Fig.3 and Table 1.

In Table 1, the row of "P2" denotes the prediction results based on the first 44 data points. In Fig.3, the prediction results are circled. Compared with the raw real data, the prediction values are wrong completely. This is because the peak value seriously affect on the following prediction values though the real values drop rapidly after the peak value.

From the above, we can see that this kind of peak value can neither be forecast nor contribute to prediction. So it should be dropped before prediction. In other words, raw data of security situation should be pre-processed before prediction, rather than direct utilization.

3 Singular spectrum analysis of time series

In fact, the original time series of network security situation is composed of two components: one is predictable component; the other is random component. We should separate the two component before prediction.

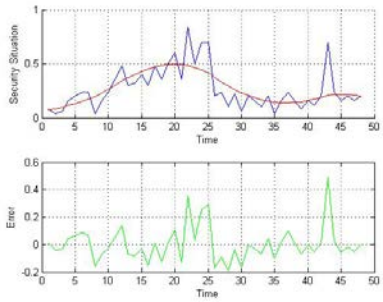

Fig.4 The result based on SSA

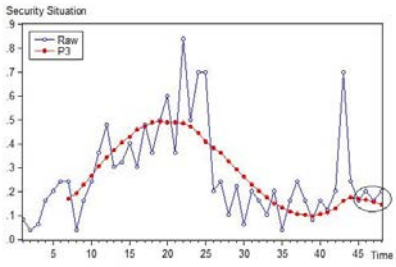

Fig.5 The prediction result based on SSA

Let X(t){x_t:1<t<N} be the time series of historical situation data. Then: $X(t)=X(t)^{(1)}+X(t)^{(2)}$, where, $X(t)^{(1)}$ denotes trend component which can be predictable, and $X(t)^{(2)}$ denotes noise component which is random completely.

We use singular spectrum analysis (SSA) to separate the two component. The SSA method is a powerful non-parametric technique of time series analysis, which belongs to the general category of Principal Component Analysis (PCA) methods.

SSA algorithm has four step: embedding, singular value decomposition, grouping, and reconstruction.

1) Embedding

The step is to map the original time series to a sequence of multidimensional lagged vectors.

Set an integer K, 1<K<N, which is called window length. Then M=N-K. So, the trajectory matrix of the time series is of dimension M×K and has the following form:

$$E=\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_k \\ x_2 & x_3 & x_4 & \cdots & x_{k+1} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ x_m & x_{m+1} & x_{m+2} & \cdots & x_{m+k} \end{pmatrix}$$

2) Singular value decomposition

The step is the singular value decomposition (SVD) of the trajectory matrix.

The matrix E's lag-covariance matrix $R=EE^T$, has M eigenvalues and eigenvectors. Let $\lambda_1, \lambda_2,..., \lambda_M$ be the eigenvalues of R ($\lambda_1 \geq ... \geq \lambda_M$). Let $U_1,...,U_M$ be the corresponding eigenvectors and $V_j=E^TU_j/\lambda_j^{1/2}$, j=1,...,M. Then the elementary matrix $E_j=\lambda_j^{1/2}U_jV_j^T$. So: $E=E_1+E_2+...+E_M$

3) Grouping

The step is to partition the set into 2 disjoint subsets $E^{(1)}$ and $E^{(2)}$ because there are two component in the time series. Then: $E^{(1)}=\Sigma_i E_i$; $E^{(2)}=E-E^{(1)}$

4) Reconstruction

The step is to group decomposition into a new series.

Compared K with M, set K*=min(K, M) and M*=max(K, M). Let e_{ij} is the elements of matrix $E^{(1)}$, then $e_{ij}^*=e_{ij}$ if K<M, otherwise $e_{ij}^*=e_{ji}$. Then

$$X(t)^{(1)}=\begin{cases} \frac{1}{t}\sum_{i=1}^{t} e^*_{i,t-i+2} & 1 \leq t < K^* \\ \frac{1}{K^*}\sum_{i=1}^{L^*} e^*_{i,t-i+2} & K^* \leq t < M^* \\ \frac{1}{N-t}\sum_{i=t-M^*+2}^{N-M^*+1} e^*_{i,t-i+2} & M^* \leq t < N \end{cases}$$

$$X(t)^{(2)}=X(t)-X(t)^{(1)}$$

So, we use the algorithm to decompose the time series in Fig.1 and get the result shown in Fig.4. The peak value of the 43rd data point has been separated successfully from main trend component. It is proven that SSA can be used to pre-process the time series of network security situation.

4 Experimental evaluation

The separated error time series, $X(t)^{(2)}$, in Fig.4 is shown that it is similar with random noise. So it should be dropped before prediction. Then, we use the first 44 data point of the time series, $X(t)^{(1)}$, to forecast the next 4 data point with the help of ARMA model.

The result is shown in Fig.5 and Table 1. In Table 1, the row of "P3" denotes the prediction results based on the first 44 data points. In Fig.5, the prediction results are circled. Although the peak value of the 43rd data point is in the time series, it does not affect on the prediction results after pre-procession.

5 Conclusion

Our analysis shows that some temporal peak values in time series of network security situation are neither forecast nor used to predicate the next value. Thus a method proposed based on SSA is to separate the random component. The experiment test can prove that the method can forecast the near future.

Acknowledgements

This work was financially supported by the Shandong Key Research and Development Program (2014GGX101021, 2015GGX101023), Shandong Natural Science Foundation (ZR2013FM025) and National Natural Science Foundation of China (61572297)

References

[1] Tim Bass, Intrusion detection systems & multisensor data fusion: creating cyberspace situational awareness, Communications of the ACM, 2000, 43(4):99-105

[2]Yu-Beng Leau, Selvakumar Manickam, Network security situation predication: a review and discussion, ICSIIT 2015, 424-435

[3]Zhang Yong, Tan Xiaobin, Xi Hongsheng, A novel approach to network security situation awareness based on multi-perspective analysis, International Conference on Computational Intelligence and Security, 2007, 768-772

[4]Yonghu Liu, Yingjie Zhou, Guangmin Hu, A backbone communication network security situation prediction method, The 2nd International Conference on Information Science and Engineering, 2010, 4215-4218

[5]Lai Jibao, Wang Huiqiang, Liu Xiaowu, Liang Ying, A quantitative prediction method of network security situation based on wavelet neural network, The 1st International Symposium on Data, Privacy and E-Commerce, 2007, 197-202

[6]Zongming Lin, Guolong Chen, Wenzhong Guo, Yanhua Liu, PSO-BPNN-based prediction of network security situation, The 3rd International Conference on Innovative Computing Information and Control, 2008,

[7]Fengli Zhang, Juan Wang, Zhiguang Qin, Using gray model for the evaluation index and forecast of network security situation, International Conference on Communications, Circuits and Systems, 2009, 309-313

[8]Dapeng Man, Yan Wang, Wu Yang, Wei Wang, A combined prediction method for network security situation, International Conference on Computational Intelligence and Software Engineering, 2010

[9]N.Nekrutkin, V.Zhigljavsky, Analysis of time series structure--SSA and related techniques, Chapman & Hall CRC, 2001, 13-7