

Using PCA to Evaluate Computer Network Security

Zhang Yunlong

Electronic Information Engineering College, Sias International University, Zhengzhou University, Xinzhen, Henan, 451150, China
Email: zhang.yunlong317@163.com

Hua Yong

Department of Foundation, The First Aeronautical College of Air Force, Xinyang, Henan, 464000, China
hy8106@163.com

Abstract—The principal components analysis (PCA) is a multivariate statistical tool which brings the multi-dimensional factor into the identical system to carry on the qualitative and quantitative research. The theory is relatively complete. This paper applies the PCA method into the synthesis evaluation of network security, thus can determine the major component which affects the network security. Relatively, this method is very practical in evaluating the network security objectively, accurately, and comprehensively, provides a new idea or method for evaluating the security status of computer network system.

Key words—network security; principal components analysis (PCA); security evaluation

I. Introduction

With the application and development of computer network technology, network security issues become increasingly prominent, network security is a growing concern. Network users must have a comprehensive understanding of the network, must have a clear understanding of network security, must have an effective measure or evaluation to the degree of network security, certainly know the security degree of their network system as well as the factors that affect the safety, in order to take appropriate preventive measures to protect the security of network systems in the maximum extent possibility. As a high-tech product, network has high complexity, uncertainty and invisibility. There are many factors that affect the security of network systems, involving many areas of computer, communications, physics, mathematics, biology, and social management, et al. Network security issues are becoming increasingly complex. Therefore, in order to obtain scientific results of an evaluation, we must comprehensively evaluate the security situation for the entire network. This paper introduces the principal component analysis (PCA) method to the evaluation of network security, elaborates the basic principle of the PCA method, studies comprehensively the integrated evaluation index system of network security, and illustrates the application process of PCA in the evaluation of network security by examples.

II. Basic Principle of the PCA Method

PCA is a dimensionality reduction method to simplify the data structure. It turns out several variables (simple indexes) into some few variables (integrated indexes). These integrated variables can reflect most of the information of the original multiple variables, and the PCA has the ability to objectively determine the weight of each index, and can avoid the

subjectivity and arbitrary of human being. In order to make these integrated variables contained in the information do not overlap each other, they should be required unrelated between themselves. Starting from these few unrelated variables, we can also get a general index which can make the problem much simpler through sorting and classifying. The PCA is a kind of multivariate analysis method, it tries to combine the many original related indexes into a new set of mutually unrelated indexes instead of the original ones. The general mathematical model of the PCA is as follows:

Suppose $X = (x_1, x_2, \dots, x_p)^T$ is a random variable with p index, get a linear combination of the indexes:

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\ \dots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p \end{cases} \quad (1)$$

Where any of the combination coefficient a_{ij} is determined by the following conditions:

$$(1) \quad a_{1j}^2 + a_{2j}^2 + \dots + a_{pj}^2 = 1, \quad j=1, 2, \dots, p$$

(2) F_1 is the one which has the largest variance in all the linear combinations of x_1, x_2, \dots, x_p ; F_2 is not related to F_1 , and its variance value is only smaller than the F_1 's in all the linear combinations of x_1, x_2, \dots, x_p ; F_p is not related to any of the F_1, F_2, \dots, F_{p-1} , and its variance is the smallest one in all the linear combinations of x_1, x_2, \dots, x_p .

Let $\Sigma > 0$ ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$) be the covariance matrix X , λ_i is non-zero characteristic root of Σ , the u_i is the standard feature vector corresponding to the λ_i . Mathematical theorems have been proved that the i -th principal component of X is $F_i = u_i^T X$ ($i = 1, 2, \dots, p$).

$\lambda_k / \sum_{i=1}^p \lambda_i$ is the contribution rate of the principal component F_k . $\sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i$ is the cumulative contribution rate of the principal components F_1, F_2, \dots, F_m .

The greater the cumulative contribution rate, the less the missing information in the data, but the amount of the subsequent calculation will increase accordingly. The standard of getting the m-Value is usually to make the cumulative

contribution rate be 85% or more.

The general steps of PCA are generally as follows:

Step 1, to select the indexes and data according to the research questions;

Step 2, to standardize the data (the Factor process of SPSS software can perform it automatically);

Step 3, to determine the correlation between the indexes;

Step 4, to determine the number of principal components;

Step 5, to determine the PCA expressions;

Step 6, to calculate the PCA values and to take the evaluation and research.

III. Building a Comprehensive Evaluation Index System

Computer network is a complex system. There are many factors affect network security in varying degrees, such as the network's external environment, the health of the server, the network transmission, the technical means of security, the organization and management system et al. To evaluate a network system whether is safe, we must consider many factors. The first job is to establish a scientific and rational network security evaluation system. The basic principles to establish a sound scientific and reasonable evaluation index system are as the following:

(1) The principle of completeness: we should take comprehensive evaluation indexes which can fully reflect the basic features of network security, so that evaluation results will be accurate and reliable.

(2) The principle of independence: that evaluation indexes tend to have a certain degree of relevance should be avoided. When the indexes are designed, their various relationships between each other should be minimized, avoiding duplication, letting the index system accurately reflect the actual situation of the network security.

(3) The principle of brief: the evaluation indexes should reflect best the level of network security; should be representative of the index system; should be structured, clear and concise.

(4) The principle of accuracy: the meaning of the evaluation indexes must be clear. If they are not clear, they will affect the evaluation results, and even let the evaluation cannot be work. All the evaluation indexes should reflect the reality of the level of network security technology.

(5) The principle of feasibility: the index system should meet the needs of the actual evaluation work, be easy of operation and evaluation, all data should be easy to collect, to facilitate computer processing, and implementation of evaluation activities to facilitate the expert.

In accordance with the above principles, the comprehensive

analysis of the impact factors the network security should begin from the analyzing of the various aspects of management security, physical security and logical security. We select 19 indexes to constitute a network security evaluation system as follows: the organization system for Security x_1 , the safety management system x_2 , the staff safety training x_3 , the emergency response mechanism, x_4 , the anti-electromagnetic leak measures x_5 , the network computer room security x_6 , the security of electricity supply x_7 , the line safety x_8 , the fault-tolerant redundant x_9 , the equipment safety x_{10} , the data backup x_{11} , the data recovery x_{12} , the system audit x_{13} , the access control x_{14} , the software safety x_{15} , the data signatures x_{16} , the anti-virus measures x_{17} , the data encryption x_{18} and the intrusion prevention x_{19} .

In the comprehensive evaluation, some of the indexes are quantitative, others are qualitative. For the quantitative indexes, their values can be gotten according to the specific circumstances of the evaluation network system. For the qualitative indexes, their values can be obtained by experts using rating or scoring evaluation method based on the specific circumstances of the evaluation system. Different indexes reflect the status of computer network security from a different angle, index values are not comparable. Therefore, we need normalize each index data in the range between 0 and 1.

IV. Application Examples

Table 1 is a network security evaluation instance, contains 10 evaluation samples, 19 quantitative data each sample, all of the data in the table have been normalized.

By using the software SPSS, we can easily calculate the correlation coefficient between the various indexes. For the instance shown in table 1, the correlation coefficients between most of the indexes are larger than 0.8. There is a strong correlation between these indexes, so they are suitable to use PCA method for processing and analysis.

Table 2 shows the analysis result of principal component extraction with the variance decomposition by using the software SPSS. The principle of determining the number of principal components is selecting the first m components which the corresponding eigenvalue is greater than 1. Eigenvalues can be to some extent as the indicators which can explain the intensity of the principal component. If the size of a eigenvalue is less than 1, it indicates that the explanatory power of the principal component is not as good as the introduction of the original average variables. Therefore, to select the principal components which corresponding eigenvalues are greater than 1 is generally as the inclusion criteria. It can be seen from Table 2, two principal components should be extracted, which corresponding eigenvalues are 16.287 and 1.069 respectively, the variance contribution rates are 85.723% and 5.628% respectively, the cumulative variance contribution rate of the two principal components of 91.351%. Therefore, the two principal components can basically reflect all of the information of the indexes.

Table 1 the Data Table

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃	X ₁₄	X ₁₅	X ₁₆	X ₁₇	X ₁₈	X ₁₉
1	1.0	0.8	0.8	0.8	1.0	0.8	0.85	0.8	0.72	0.8	0.92	0.87	0.85	0.8	0.93	1.0	0.8	0.8	0.9
2	1.0	1.0	0.8	0.8	1.0	0.9	0.85	0.8	0.77	0.8	0.93	0.90	0.90	1.0	0.95	1.0	0.8	0.6	0.9
3	0.8	0.8	0.6	0.6	1.0	0.8	0.80	0.8	0.78	0.8	0.89	0.85	0.90	0.8	0.91	1.0	0.8	1.0	0.8
4	0.8	0.8	0.6	0.6	0.0	0.6	0.65	0.6	0.67	0.6	0.81	0.72	0.75	0.6	0.82	0.0	0.8	0.4	0.7
5	0.6	0.8	0.6	0.6	0.0	0.5	0.65	0.6	0.63	0.6	0.77	0.75	0.70	0.6	0.81	0.0	0.6	0.4	0.7
6	0.4	0.6	0.4	0.4	0.0	0.4	0.45	0.4	0.52	0.4	0.65	0.63	0.70	0.6	0.67	0.0	0.6	0.4	0.6
7	0.6	0.4	0.2	0.2	0.0	0.5	0.50	0.4	0.61	0.6	0.62	0.61	0.65	0.4	0.55	1.0	0.6	0.2	0.6
8	0.4	0.4	0.2	0.2	0.0	0.3	0.25	0.4	0.43	0.4	0.47	0.55	0.45	0.4	0.47	0.0	0.4	0.2	0.4
9	0.2	0.4	0.2	0.2	0.0	0.3	0.30	0.2	0.27	0.2	0.33	0.37	0.25	0.2	0.35	0.0	0.2	0.2	0.4
10	0.8	0.8	0.4	0.6	1.0	0.5	0.60	0.6	0.64	0.6	0.75	0.77	0.80	0.6	0.78	0.0	.06	0.4	0.7

Table 2 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	16.287	85.723	85.723	16.287	85.723	85.723
2	1.069	5.628	91.351	1.069	5.628	91.351
3	.627	3.299	94.649			
4	.355	1.866	96.515			
5	.302	1.590	98.106			
6	.138	.724	98.830			
7	.096	.503	99.333			
8	.070	.370	99.703			
9	.056	.297	100.000			
10	3.67E-016	1.93E-015	100.000			
11	2.77E-016	1.46E-015	100.000			
12	1.55E-016	8.16E-016	100.000			
13	8.48E-017	4.46E-016	100.000			
14	2.68E-017	1.41E-016	100.000			
15	-6.69E-017	-3.52E-016	100.000			
16	-1.33E-016	-6.99E-016	100.000			
17	-2.69E-016	-1.42E-015	100.000			
18	-4.92E-016	-2.59E-015	100.000			
19	-2.96E-015	-1.56E-014	100.000			

Using the data in Table 3 (the principal component loading matrix) to be divided by the eigenvalue corresponding to the principal component, and then their square root, the coefficients corresponding to the principal components can be obtained. Thus, we can obtain the principal components expressions as follows:

Table 3 the Initial Factor Loadings Component Matrix

	Component			Component	
	1	2		1	2
X ₁	.235	.035	X ₁₁	.244	-.085
X ₂	.223	-.367	X ₁₂	.243	-.058
X ₃	.227	-.223	X ₁₃	.235	-.030
X ₄	.230	-.286	X ₁₄	.236	-.063
X ₅	.194	.209	X ₁₅	.242	-.182
X ₆	.236	.194	X ₁₆	.254	.730
X ₇	.242	.017	X ₁₇	.230	-.018
X ₈	.243	-.007	X ₁₈	.207	.176
X ₉	.237	.058	X ₁₉	.244	-.003
X ₁₀	.237	.189			

$$F_1 = 0.235x_1 + 0.223x_2 + 0.227x_3 + 0.230x_4 + 0.194x_5 + 0.236x_6 + 0.242x_7 + 0.243x_8 + 0.237x_9 + 0.237x_{10} + 0.244x_{11} + 0.243x_{12} + 0.235x_{13} + 0.236x_{14} + 0.242x_{15} + 0.154x_{16} + 0.230x_{17} + 0.207x_{18} + 0.244x_{19}$$

$$F_2 = 0.035x_1 - 0.367x_2 - 0.223x_3 - 0.286x_4 + 0.209x_5 + 0.194x_6 + 0.017x_7 - 0.007x_8 + 0.058x_9 + 0.189x_{10} - 0.085x_{11} - 0.058x_{12} - 0.030x_{13} - 0.063x_{14} - 0.182x_{15} + 0.730x_{16} - 0.018x_{17} + 0.176x_{18} - 0.003x_{19}$$

$$8x_{12} - 0.030x_{13} - 0.063x_{14} - 0.182x_{15} + 0.730x_{16} - 0.018x_{17} + 0.176x_{18} - 0.003x_{19}$$

Where zx_i is the standardized value of x_i , $i=1,2,\dots,19$. Take the standardized values in Table 1 into the above formula, we can get the values of the principal component corresponding to each sample. Sorting them, we can be obtained the evaluation order of each department depending on the PCA. Need to explain, if the numbers of eigenvalues which are greater than 1 are more than one, sometimes two or even three or four. This time, we can take their variance contribution rates as the respective weights, being multiplied by their corresponding principal components values, add up to get a comprehensive evaluation expression. Or, we can take their variance contribution rates as the respective weights, being multiplied by their corresponding coefficients corresponding to the principal components, add up to get a series of comprehensive coefficients, and finally get a principal component model as follows:

$$F = 0.223x_1 + 0.187x_2 + 0.199x_3 + 0.198x_4 + 0.194x_5 + 0.234x_6 + 0.228x_7 + 0.227x_8 + 0.226x_9 + 0.234x_{10} + 0.224x_{11} + 0.225x_{12} + 0.219x_{13} + 0.217x_{14} + 0.216x_{15} + 0.190x_{16} + 0.215x_{17} + 0.205x_{18} + 0.229x_{19}$$

Using this principal component model, we can calculate the value of the integrated principal component, sorting them, we can comprehensively evaluate and compare the network security status of each sample, the results are shown in Table 4.

Table 4 the Score of Principle Components

Samples	The Score of Principle Components	Order
1	3.491	2
2	3.590	1
3	3.383	3
4	2.512	5
5	2.337	6
6	1.922	8
7	2.009	7
8	1.386	9
9	0.997	10
10	2.582	4

V. Conclusions

This paper introduces the PCA method of the multivariate statistical analysis into the comprehensive evaluation of network security. The PCA method can avoid the loss of the

characteristics information of each index variation, can more accurately reflect all of the information contained in the data, can reduce the workload of the modeling effectively.

References

- [1] Xiao DJ, Yang SJ, Zhou KF, Chen XS. An evaluation model of network security. Journal of Huazhong University of Science and Technology(Nature Science Edition),2002,30(4):37-39(in Chinese with English abstract).
- [2] Kayashima M, Nagai Y, Terada M. Network Security for the Broadband Era[J].Hitachi Review,2002,51(2):70-73.
- [3] Stephen S Yau, Xinyu Zhang. Computer Network Intrusion Detection, Assessment and Prevention Based on Security Dependency Relation[C]. In: IEEE 23RD Annual International Computer Software & Applications Conference,1999:86~91.
- [4] Jolliffe I.T, Principal Component Analysis [M], New York:Springer,1986.
- [5] Ma Hong. The Application of principal ingredient analysis method to the water quality appraisalment[J]. Journal of Nanchang Institute of Technology,2006,25(1):65-67.