Network Anomaly Detection Method Based on I-KPCA

Jiandong Shang[1, a], Qiang Li[2, b], Runjie Liu[3, c], Yuting Niu[4, d]

[1]School of Zhengzhou's Smarter City University, Henan 450000, China;

[2] School of Zhengzhou's Smarter City University, Henan450000, China;

[3] School of Zhengzhou's Smarter City University, Henan450000, China;

[4] School of Zhengzhou's Smarter City University, Henan450000, China.
[a]shangjiandong@zzu.edu.cn, [b]LiQiang18438570019@163.com, [c]ierjliu@zzu.edu.cn,
[d]nyt37977221@163.com

Abstract. Network anomaly detection is a hot topic in the field of detection and is of great significance for ensuring the reliable operation of the network. The current research direction is mainly the detection of the host's own operating conditions, and the detection of a single resource, low detection efficiency cannot meet the real-time detection needs and other issues. Based on the theory of Kernel Principal Component Analysis (KPCA), this paper proposes an improved I-KPCA network anomaly detection method, which can integrate multiple data resources for evaluation and greatly reduce the false alarm rate. In order to verify the performance of the detection method, this article focuses on comparative experiments conducted in the Matlab environment. The experimental results show that the network anomaly detection method based on the improved KPCA can not only detect the abnormal situation in real time, but also make the false alarm rate not exceed 0.85% and the detection rate reaches 96%.

Keywords: I-KPCA, abnormal detection, Nuclear principal component analysis, Guass-Seidel.

1. Introduction

Nowadays, all aspects related to the Internet and the Internet have developed rapidly. In particular, the development of services on the Internet platform is more rapid, and people obtain the services they need through various channels on the Internet platform. However, the ensuing problem is that people are gradually focusing their attention on the quality and effectiveness of the services provided by the network platform. The rapid detection of the server operation and the effective notification of the results of the anomaly detection have become an important problem at present.

With regard to the anomaly detection method of the network, a lot of research achievements have been made in academic circles at home and abroad. In 2002, BARFORDP et al. introduced wavelet analysis into network anomaly detection, used wavelet filtering network anomaly flow, analyzed network anomalous flow after the experiment, and judged the anomaly based on the change of data after the experiment. In 2008, EAMONN KEOGH [1] proposed the SAX method based on time series search. This method judges network abnormalities by comparing time series inconsistency events. In 2012, Ming Chen and Lixin Ye [2] proposed an MSPCA-based full-network anomaly detection method. This method uses the multi-scale modeling capability of wavelet transform and the dimensionality reduction capability of PCA to model the normal flow, and then uses The SHEWART control chart and the EWMA control chart analyze the residual flow.

The main research direction of this paper lies in the abnormality of the running state of the server itself. Through the multi-dimensional monitoring of the server state, the abnormal situation of the network itself is discovered. The research focuses on the monitoring of the hosts of the high-speed network on the service platform, starting with the collection of various data indicators of the host's own state on the service platform, and then using the improved core principal component analysis (I-KPCA) method to collect the collected data. The data was model-analyzed, and correlation experiments were performed on Matlab software. The performance of the model was judged by its detection rate and false alarm rate.

2. Establishment of I-KPCA Model

2.1 KPCA Implementation Steps

The KPCA analysis method is based on the spatial sample and has nothing to do with the dimensions of the original spatial data input. The complexity is proportional to the number of samples. Since the collected data is multidimensional data, the KPCA [3] analysis method is used to solve the matrix's covariance, and the eigenvalues and feature vectors of the covariance matrix are obtained. We form new variables through certain mathematical combinations. We call these variables new principal elements. Select the pivot element that can represent most of the matrix information as a new sample of the experiment.

1.Obtaining k data indicators from network resources, each of which contains l attribute, writes the obtained data into a $k \times l$ -dimensional matrix, which is expressed as follows:

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1l} \\ \vdots & \ddots & \vdots \\ b_{k1} & \cdots & b_{kl} \end{bmatrix}$$

2. Select the RBF kernel function and calculate the N matrix above. The expression of the RBF function is as follows:

$$K(x, x_j) = \exp\left(-\frac{\left\| x - x_j \right\|^2}{\sigma^2} \right)$$

3. According to the nuclear matrix N, it is corrected and NL is obtained.

4. Calculate the eigenvalue $\lambda_i(1,2,\cdots,k)$ and eigenvector $v_j(1,2,\cdots,k)$ of NL using the Gauss-Seidel method.

5. The required $\lambda_i(1,2,\cdots,k)$ is arranged in the bottom row, and the corresponding $v_j(1,2,\cdots,k)$ changes along with the change.

6. Use the $Gram-Schmidt$ orthogonal method to unitize the requested $v_j(1,2,\cdots,k)$ to get $\beta_1, \beta_2, \cdots, \beta_k$.

7. Calculate $\lambda_i(1,2,\cdots,k)$ cumulative contribution rate $B_1, B_2, \cdots, B_k$, according to the given extraction rate p, if $B_t \geq p$, then extract t principal components $\beta_1, \beta_2, \cdots, \beta_t$.

8. The modified matrix NL is calculated, and the extracted feature vector is projected, that is, the reduced dimension data. That is $Y = NL \cdot \beta$, where $\beta = (\beta_1, \beta_2, \cdots, \beta_k)$.

2.2 Using I-KPCA Abnormal Detection

The key to diagnose using the I-KPCA method in the network abnormality monitoring system is to find a suitable kernel function and determine whether the network is abnormal by determining the size of the statistic contribution rate T^2. However, it is worth noting that what T^2 represents is the range fluctuation of one of the variables in the sample. If the variable has little relationship with the sample pivot element, the change in the variable cannot be represented by T^2-graph. Based on this situation, we must determine whether the network is abnormal by detecting the Q statistic.

First, find the original function according to the I-KPCA solution steps above. Here we use the RBF kernel function:

$$l(x_j, x_l) = l(u \cdot x_j, u \cdot x_l) = \exp\left(\frac{\left\| u \cdot x_j - u \cdot x_l \right\|^2}{\sigma} \right) \tag{1}$$

Second, find $\frac{\partial u}{\partial u_i}$ for the i variable in the above formula u:

$$\frac{\partial l(x_j, x_l)}{\partial u_i} = \frac{\partial l(u \cdot x_j, u \cdot x_l)}{\partial u_i} = -\frac{1}{\sigma}\left(u_i x_{j,i} - u_i x_{l,i}\right)^2 l(u \cdot x_j, u \cdot x_l) = -\frac{1}{\sigma}\left(x_{j,i} - x_{l,i}\right)^2 l(x_j, x_l)\big|_{u_i = 1} \tag{2}$$

It can be seen from the above that the size of the influence kernel function depends on the absolute value of the partial derivative of the variable in the sample.

Similarly, we can find the partial derivative of the product of the corresponding kernel function based on this method:

$$\frac{\partial l(x_j, x_l)}{\partial u_i} = \frac{\partial l(x_j, x_N)}{\partial u_i} = -\frac{1}{\sigma}\left[(x_{j,i} - x_{N,i})^2 + (x_{l,i} - x_{N,i})^2\right] \bullet l(x_j, x_N) \cdot l(x_l, x_N) \tag{3}$$

the above formula x_N is the re-acquired sample data.

from the above I-KPCA principle, statistics T^2 can be written as:

$$T_N^2 = t_N^{\mathrm{T}} \Lambda^{-1} t_N = \overline{L_N^T} \alpha \Lambda^{-1} \alpha^{\mathrm{T}} \overline{L_N} = tr\left(\alpha^{\mathrm{T}} \overline{L_N}\, \overline{L_N^T} \alpha \Lambda^{-1}\right) \tag{4}$$

formula (4) can be obtained from Formula (5):

$$C_{T^2,N,i} = \left|\frac{\partial T_N^2}{\partial u_i}\right| = \left|\frac{\partial}{\partial u_i}\left(tr\left(\alpha^{\mathrm{T}} \overline{L_N}\, \overline{L_N}^{\mathrm{T}} \alpha \Lambda^{-1}\right)\right)\right| = \left|tr\left(\alpha^{\mathrm{T}}\left(\frac{\partial}{\partial u_i} \overline{L_N}\, \overline{L_N}^{\mathrm{T}}\right)\alpha \Lambda^{-1}\right)\right| \tag{5}$$

From Equations (3)-(5), it can be seen that to obtain the contribution rate of T^2, we need to know the value of $\overline{L_N}\,\overline{L_N}^{\mathrm{T}}$ first, and then push out the value of $\frac{\partial\left(\overline{L_N}\,\overline{L_N}^{\mathrm{T}}\right)}{\partial u_i}$. Finally, we can substitute these two values into (5) to get a new sample. The $C_{T^2,N,i}$ in each variable then draws a corresponding graph to determine if the network has anomaly.

3. Analysis of Results

In this paper, Matlab (R2014a) software is used as the experimental environment to simulate the proposed I-KPCA detection method. The data sources used were samples collected by multiple servers in the laboratory. The frequency of collection was collected every 30 seconds for a total of 60 acquisitions. Because the experimental environment for collecting sample data is unstable, the experimental model established can only reflect the situation within a short time.

In order to evaluate and compare the performance of the two network anomaly detection methods, this paper uses two performance indicators: detection rate and false alarm rate. The detection rate refers to the ratio of the number of abnormal data samples detected by the algorithm to the total number of actual abnormal data samples. The false alarm rate refers to the ratio of the number of normal data samples that are misjudged by the algorithm as abnormal to the total number of normal data samples.

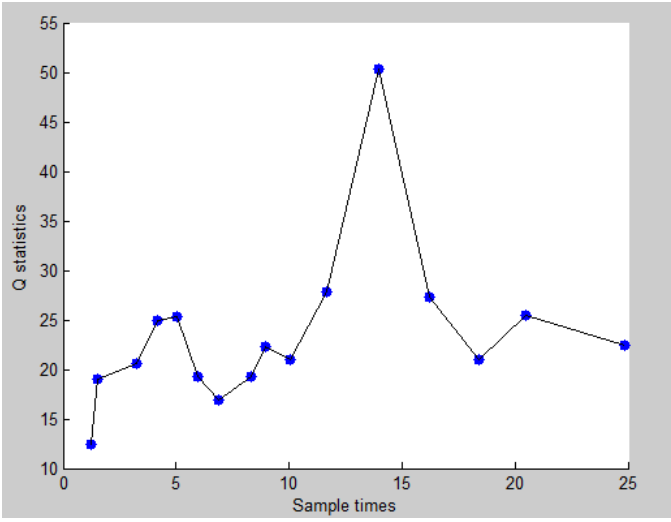

Fig 1. Q statistics graph

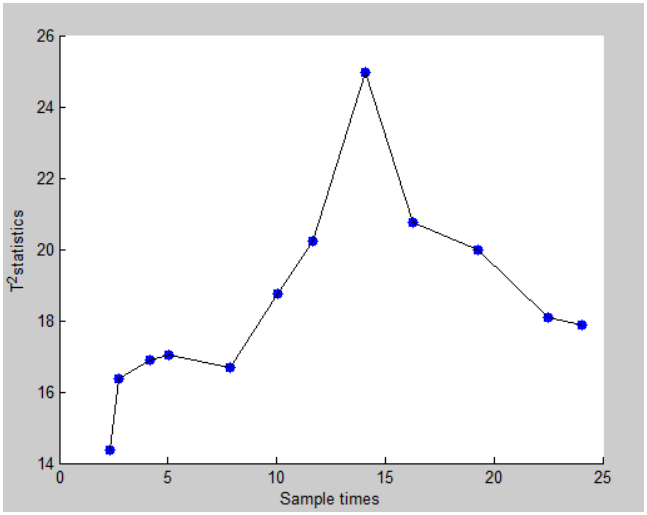

Fig 2. T^2 statistics graph

As these two statistics are theoretically speaking there is no so-called upper bound. When the changes of these two variables are large, we can think that the sample has a greater impact, that is, the network fails. As can be seen from the above figure, an abnormal situation occurred on the 13th network.

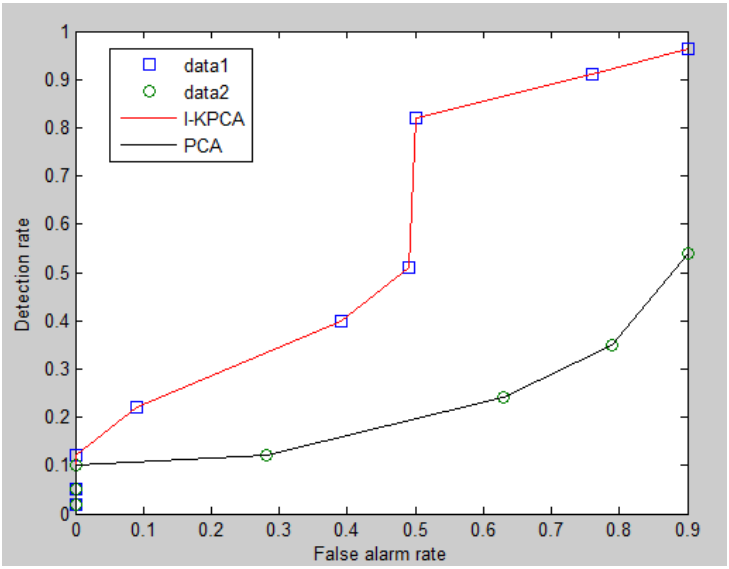

Fig 3. Comparison of detection rate and false alarm rate

The experimental results show that the abnormal detection method based on I-KPCA can not only detect the abnormal situation in real time, but also make the false alarm rate not exceed 0.85% and the detection rate reaches 96% compared with the traditional PCA method.

4. Conclusion

This paper proposes an I-KPCA-based method for rapid detection of abnormalities. Gauss-Seidel method is used to reduce the dimension of acquired data, and new samples are re-determined. The kernel function is determined from the extracted samples. And established the I-KPCA model by analyzing the statistical contribution rate of sample variables to determine whether the network is abnormal. Through the test rate and the false alarm rate, the experimental evaluation was carried out. It can be known from the experiment that the detection rate based on the improved KPCA method is greatly improved.

References

[1]. Shen Lin, Research and application of time series anomaly detection. Hohai University. 2008, 40.

[2]. Yekui Qian, Ming Chen, Lixin Ye. Anomaly Detection Method Based on Multi-scale Principal Component Analysis in All Networks[J]. Journal of Software, 2012(2): 1000-9825.

[3]. Choi, SW. Lee, et al. Fault detection and identification of nonlinear processes based on kernel PCA CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, JAN 28 2005,p55-67.

[4]. Lee, JM; Yoo, CK; Choi, SW; et al. Nonlinear process monitoring using kernel principal component analysis CHEMICAL ENCINEERING SCIENCE. JAN2014, p223-234.

[5]. Brauckhoff D, Salamatian K, May M. Applying PCA for traffic anomaly detection: problems and solutions. In: Proc. of the INFOCOM. New York: IEEE Press, 2009. 46–53.