Anomaly Detection in Big Data with Separable Compressive Sensing

Wei Wang, Dan Wang, Shu Jiang, Shan Qin and Lei Xue

Abstract Nowadays, the anomaly detection of big data is a key problem. In this setting, principal components analysis (PCA) as an anomaly detection method is proposed, but PCA also has scalability limitations. Thus, we proposed the feasibility measure to use the PCA and separable compression sensing to detect the abnormal data. Subsequently, we prove that volume anomaly detection using compressing data can achieve equivalent performance as it does using the original uncompressed and reduces the computational cost significantly.

Keywords Big data • Anomaly detection • Separable compression sensing • PCA

1 Introduction

The arrival of the era of big data promoted the development of information retrieval and data mining technology [1]. Detection of volume abnormal information is also becoming more and more important. There are a lot of detecting problem of large data in many practical applications. Furthemore, the exception will make network congestion and will cause serious influence to the user, thus analysis of abnormal problem is very important for us [2].

In recent work demonstrated a useful role for principal component analysis (PCA) to detect network anomalies. They showed that the minor components of PCA (the subspace obtained after removing the components with largest eigenvalues) revealed anomalies that were not detectable in any single node-level trace. This work assumed an environment in which all the data is continuously pushed to

W. Wang (✉) · D. Wang · S. Qin · L. Xue
College of Electronic and Communication Engineering, Tianjin, China
e-mail: weiwangvip@163.com

S. Jiang
Research Institute, China Shipbuilding Industry Corporation, No.704, Shanghai, China

© Springer-Verlag Berlin Heidelberg 2016

Q. Liang et al. (eds.), *Proceedings of the 2015 International Conference on Communications, Signal Processing, and Systems*, Lecture Notes in Electrical Engineering 386, DOI 10.1007/978-3-662-49831-6_59

a centralsite for off-line analysis. Such a solution cannot scale either for networks with a large number of monitors nor for networks seeking to track and detect anomalies at very small time scales. Thus, anomaly detection in large data is still a problem to be studied.

In this paper we propose a general method to diagnose anomalies. This method is based on PCA (principal components analysis) algorithm and the CS (compression sensing) theory to realize the data anomaly detection. The goal of this article is in order to achieve the data of anomaly detection.

The paper is organized as follows. In Sect. 1, we introduced the data of the research status of anomaly detection and research content. In Sect. 2, we introduced anomaly detection theory and separable compression sensing theory. In Sect. 3, we first generate simulation data and then the data for training and testing results are obtained. In Sect. 4, we get the article conclusion.

2 Theory

2.1 Anomaly Detection

According to Lakhina et al. [2], we can learn a lot of anomalies which is very rare, and abnormal will be hidden in the normal data. PCA [3] is a small number of principal components derived from the original variables, so that they can retain the information of the original variables as much as possible [4]. Thus, we can use PCA algorithm which is easier to find abnormal.

First, let the network information by matrix $X = (X_1, X_2, \ldots, X_L)$ and each of these data $X_i \in R^N, i = 1, \ldots, L$ [5] after the normalization of matrix and we begin to decomposition for the normalized matrix by using PCA algorithm, then get the same covariance matrix:

$$\Sigma x = \left(\frac{1}{L}\right) XX^T = U\Lambda U^T \tag{1}$$

Given that $U = [U_1, \ldots, U_k]$ are the principal eigenvectors of $\sum_x$ corresponding to the largest K eigenvalues $\lambda_1, \ldots, \lambda_K$, the projection onto the residual subspace is $P = (I - UU^T)$ [6]. So, for the checked data X, its protection into the residual subspace is $Z = PX = (I - UU^T)X$. If Z follows a multivariate normal distribution, the squared prediction error (SPE) [7] statistic is given as:

$$tSPE = \|Z\|_2^2 = \left\|(I - UU^T)X\right\|_2^2 \tag{2}$$

and follows a noncentral chi-square distribution under the null hypothesis that the data is ‘normal’. Hence, rejection of the null hypothesis can be based on whether tSPE exceeds a certain threshold corresponding to a desired false alarm rate β. In [2], the Q-statistic was identified as threshold, it is usually expressed as:

$$Q_{\beta} = \theta_1 \, [\frac{c_{\beta}\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2}]^{\frac{1}{h_0}} \tag{3}$$

where $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $\theta_i = \sum_{j=K+1}^{N} \xi_j^i$ for $i = 1, 2, 3$, $c_{\beta} = (1 - \beta)$ percentile in a standard normal distribution and Q_{β}, and ξ_j, $i = 1, \ldots, M$ are the eigenvalues of $\sum_y$.

Once the tspe > Q_{β}, then can estimate the data which is the abnormal state detection data.

2.2 Theory of Separable Compression Sensing

Separate compression sensing theory pointed out that the random measurement matrix through the tensor product can be expressed as $\Phi = \Phi_x \otimes \Phi_y$ [8], Depending on the theory of CS X Random projection value G can be represented as:

$$G = \Phi_x X \Phi_y \tag{4}$$

The (1) further the available data X of the sparse coefficient matrix is expressed as A:

$$G = \Phi_x \Psi_x A (\Psi_y)^T \Phi_y \tag{5}$$

The $\Psi = \Psi_x \otimes \Psi_y$ is sparse transformation matrix tensor product form. Depending on the random measured value, can reconstruct the original data:

$$\min \|A\|_1 \qquad \text{subject to}\, G = \Phi_x \Psi_x A (\Psi_y)^T \Phi_y \tag{6}$$

According to the Eq. (5), G represents the compressed of testing data, the G instead of type (2) in X, then carries on the data of the anomaly detection [9, 10].

3 Simulation and Experimental Results

3.1 Synthetic Data

There are two main purpose of the experiment. The first one is to prove that our method is feasible for Large N. The second is to show that it has a better time resolution property when processing the data stream in CS domain. First, we consider a model for the network data as follows:

$$x = s + n \tag{7}$$

where x is the snapshot of network traffic over N links: Which letter s represents original signal and the letter n is a gaussian noise that is added to the original data [6]. We selected matrix which is 500 lines of 2000 columns. In order to make the anomaly in the compression state as much as possible is detected, we selected the multiple compression matrix for the test, at the same time, we add the Gaussian noise average is zero-mean and mean square error is 0.01. To imitate the network abnormal, we added 40 sample sizes of exception information on the original signal following the procedure mentioned in [2].

3.2 Experiments Results

We investigated the uncompressed matrix and the compression degree of different matrix analysis as shown in the results. Horizontal axis represents the β from 0.1 to 0.9, ordinate represents anomaly detection accuracy. Specifically, Fig. 1 shows a plot of the eigenvalue distribution between original and compressed data. This is a

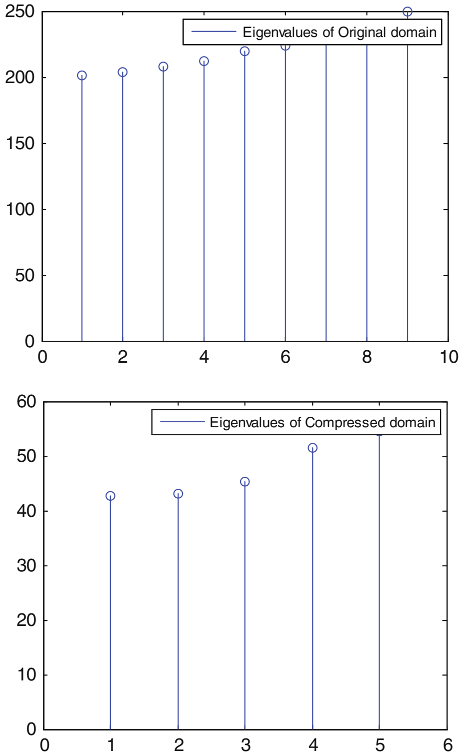
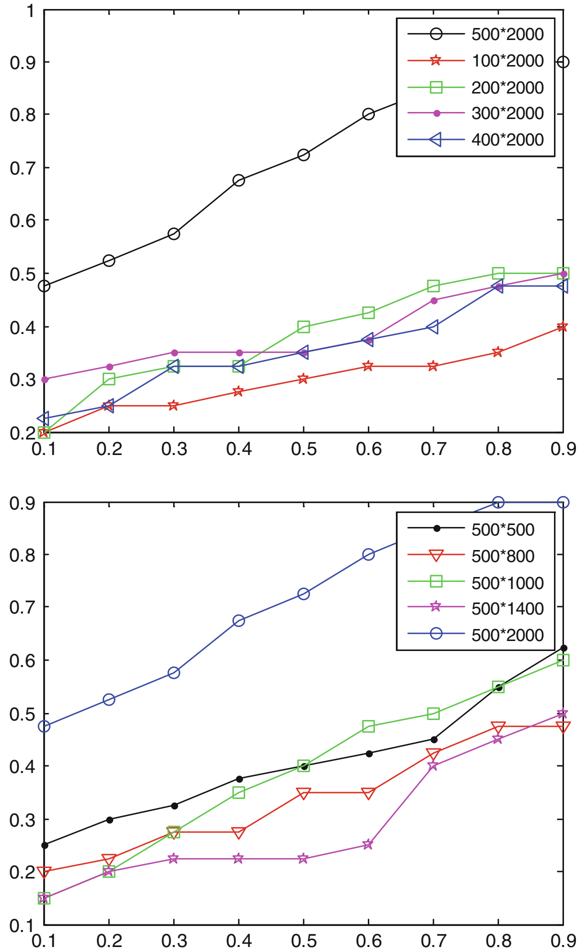


Fig. 1 Eigenvalues of original and compressed data

Fig. 2 The correct probability of anomaly detection with different matrix



very encouraging result from the point of view of detecting anomaly in CS domain. Under each beta we all joined the anomaly information with 40, the number of each matrix to detect abnormal is $Y/40$ and the anomaly detection accuracy can be calculated according to Fig. 2. Uncompressed curve as shown in Fig. 2, it can detect the abnormal information ratio which increases with the beta. As shown in Fig. 2 matrix compression degree can detect abnormal probability which is also different. The smaller the compression degree of anomaly detection the accuracy is higher.

4 Conclusions

Through simulation experiment, we use PCA and separable compression sensing to detect the different matrices, the matrix of uncompression is more easily to detect the abnormal than the matrix of compression. Thus, we have to choose the degree of compression in order to detect the abnormal information more accurately.

Acknowledgments This paper is supported by Natural Science Foundation of China (61271411), and cooperation project of 2014 annual national natural science fund committee with the Edinburgh royal society of British (613111215). It also supported by Tianjin Research Program of Application Foundation and Advanced Technology (15JCZDJC31500), National Youth Fund (61501326) and Research on real time topology optimization and efficient broadcasting transmission algorithm in ZigBee networks for smart grid (61401310).

References

1. Zhang J, Li H, Gao Q, Wang H, Luo Y (2014) Detecting anomalies from big network traffic data using an adaptive detection approach. J Inf Sci 318:91–110
2. Lakhina A, Crovella M, Diot C (2004) Diagonising network-wide traffic anomalies. In: Proceedings of the ACM SIGCOMM
3. Wang M (2007) A method for detecting wide-scale network traffic anomalies. J ZTE Commun 4:19–23, 1671-5799
4. Kanda Y, Fontugne R, Fukuda K, Sugawara T (2013) Anomaly detection method using entropy-based PCA with three-step sketches. J Comput Commun 36(5):575–588
5. Pham D-S, Venkatesh S, Lazarescu M, Budhaditya S (2014) Anomaly detection in large-scale data stream networks. J Data Min Knowl Disc 28:145–189. doi:10.1007/s10618-012-0297-3
6. Pham DS, Saha B, Lazarescu M, Venkates S (2009) Scalable network-wide anomaly detection using compressed data, Perth, W.A.
7. Ling H (2006) In-network PCA and anomaly detection. In: Proceedings of the twentieth annual conference on neural information processing systems 19, Vancouver, British Columbia, Canada, 4–7 Dec 2006
8. Rivenson Y (2009) Practical compressive sensing of large images. In: 2009 IEEE 16th international conference on digital signal processing. doi:10.1109/ICDSP.2009.5201205
9. Rivenson Y, Stern A (2009) Compressed imaging with a separable sensing operator. IEEE Signal Process Lett. doi:10.1109/LSP.2009.2017817
10. Wang W, Dunqiang L, Zhou X, Zhang B, Jiasong M (2013) Statistical wavelet-based anomaly detection in big data with compressive sensing. EURASIP J Wireless Commun. Networking. doi:10.1186/1687-1499-2013-269