Network Anomaly Detection and Visualization using Combined PCA and Adaptive Filtering

Altyeb Altaher, Sureswaran Ramadass
NAv6 Center of Excellence, Universiti Sains Malaysia
USM, 11800 Penang, Malaysia
altyeb@nav6.usm.my,sures@nav6.usm.my,

Noureldien Abdelrahman , Ahmed Khalid
Faculty of Computer Sciences and IT, University of Sciences and Technology, Sudan
noureldien@hotmail.com ,asalih2@hotmail.com

Abstract

In recent years network anomaly detection has become an important area for both commercial interests as well as academic research. This paper provides a Combined Principal Component Analysis (PCA) and Filtering Technique for efficient and effective detection and identification of network anomalies. The proposed technique consists of two stages to detect anomalies with high accuracy. First, we apply the Principal Components Analysis to transform the data to a new coordinate system such that the projection on the coordinate contains the greatest variance. Second, we filter traffic to separate between the normal and anomalous traffic using adaptive threshold. Our analysis results from network-wide traffic datasets show that our proposed provides high detection rate, with the added advantage of lower complexity

Keywords- Network anomaly detection, principal component analysis , network anomaly visualization, adaptive network traffic filter.

I. INTRODUCTION

Detecting unexpected changes in traffic patterns is a topic which has recently received much attention from the network measurement community. Network traffic is often seen to exhibit sudden deviations from normal behavior. Some of these deviations are caused by malicious network attacks such as Denial-Of-Service or viruses, whereas others are the result of equipment failures and accidental outages [1].Heady et al.[8] defined an intrusion as "any set of actions that attempt to compromise the integrity , confidentiality or availability of information resources". The identification of such a set of malicious actions is called intrusion detection problem that has received great interest from researchers. Several schemes proposed in the literature are derived from classical time series forecasting and outlier analysis methods and applied to the detection of anomalies or faults in networks [9, 10, 11].

Principal Component Analysis [3] (PCA) is a good statistical-analysis technique for detecting network traffic anomalies. PCA is used to separate the high-dimensional space occupied by a set of network traffic measurements into two disjoint subspaces corresponding to normal and anomalous network conditions. The main advantage of this approach is that it exploits correlations across links to detect network- wide anomalies. Recent papers in networking literature have applied PCA to the problem of traffic anomaly detection with promising initial results [4, 2, 5, 1].

The proposed approach consists of two stages to detect anomalies with high accuracy. First, we apply the Principal Components Analysis to transform the data to a new coordinate system such that the projection on the coordinate contains the greatest variance. Second, we filter traffic to separate between the normal and anomalous traffic using adaptive threshold.

II . Combined PCA and Adaptive Filtering Approach

This section presents the proposed approach. In Section A, we describe the PCA based intrusion detection that is utilized for detecting the anomaly traffic. In section B we describe the Combined PCA and Adaptive Filtering Approach.

A. Principal Component Analysis

Principal Component Analysis (PCA, also called Karhunen-Loeve transform) is one of the most widely used dimensionality reduction techniques for data analysis and compression. It is based on transforming a relatively large number of variables into a smaller number of uncorrelated variables by finding a few orthogonal linear combinations of the original variables with the largest variance. The first principal component of the transformation is the linear combination of the original variables with the largest variance; the second principal component is the linear combination of the original variables with the second largest variance and orthogonal to the first principal component and so on. In many data sets, the first several principal components contribute most of the variance in the original data set, so that the rest can be disregarded with minimal loss of the variance for dimension reduction of the data [6, 7]. The transformation works as follows.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Given a set of observations *x*1, *x*2, . . . , *xn*, where each observation is represented by a vector of length m, the data set is thus represented by a window X *n*×*m*

$$X_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}=\left[x_1, x_2, \cdots, x_n\right] \quad (1)$$

The average observation is defined as

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i \quad (2)$$

The deviation from the average is defined as

$$\Phi_i = x_i - \mu \quad (3)$$

The sample covariance matrix of the data set is defined as

$$C=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T=\frac{1}{n}\sum_{i=1}^{n}\Phi_i\Phi_i{}^T=\frac{1}{n}AA^T \quad (4)$$

Where $A=\left[\Phi_1, \Phi_2, \cdots, \Phi_n\right]$

To apply PCA to reduce high dimensional data, eigenvalues and corresponding eigenvectors of the sample covariance matrix C are computed. We choose the k eigenvectors having the largest eigenvalues. Often there will be just a few large eigenvalues, and this implies that k is the inherent dimensionality of the subspace governing the ”signal” while the remaining (m - k) dimensions generally contain noise [7].

We form a m x k matrix U whose columns consist of the k eigenvectors. The representation of the data by principal components consists of projecting the data onto the k-dimensional subspace according to the following rules [7]

$$y_i = U^T(x_i-\mu)=U^T\Phi_i \quad (6)$$

B. The Proposed Detection Approach

Principal component analysis has been applied to the intrusion detection as a data reduction technique not as an anomaly identifier .In this paper we combine the PCA with adaptive filter to identify anomalies in network traffic.

Based on statistical analysis, we assume that the used data set has a normal distribution, we propose suitable analysis and detection techniques to detect anomalies with high confidence while reducing the false acceptances

The proposed Combined PCA and Adaptive Filtering Approach consist of the following steps:

1) Fix the window size equal to N and (In our simulations we used N= 41)

2) Apply the PCA in network traffic window to identify patterns in network traffic, and express the network traffic in such a way as to highlight their similarities and differences.

3) Calculate the mean and the standard deviation of network traffic window

4) if the network traffic in the window exceeds the threshold Ω it considered as anomaly.

The threshold Ω defined as follow

$$\Omega=\mu+c\sigma \quad (7)$$

Where c = 2.25

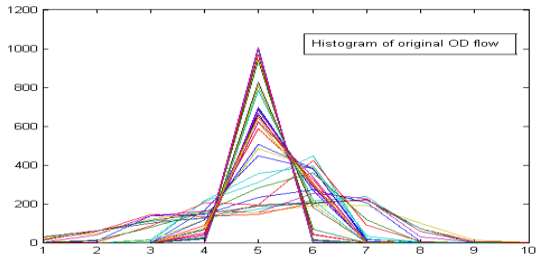
III. EXPERIMENTS

A. Data

We used the Abilene dataset, this dataset was collected from 11 core routers in the Abilene backbone network for a week (Dec. 15 to Dec. 21, 2003). It comprises two multivariate time series , one being the number of packets and the other the number of individual IP flows in each of the Abilene backbone flows (the traffic entering at one core router and exiting at another), binned at five minute intervals. Both datasets, X(1) and X(2), are of dimension F × T , where T = 2016 is the number of time steps and F = 121 is the number of backbone flows[2].

B.Anomaly Detection using the combined PCA and Adaptive Filtering

To gain a clearer understanding of the nature of the Abilene data set , we examine the Histogram of the Abilene data set as in Fig 1.the shape of histogram indicates that data is normally distributed , as a normal distribution is characterized by its bell shape. The curve of histogram is concentrated in the center and decreases on either side, this means that the data set has less of a tendency to produce unusually extreme values. Fig 2 is a plot of Abilene data set .



Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig 1: Histogram of original OD flow from Abilene data set

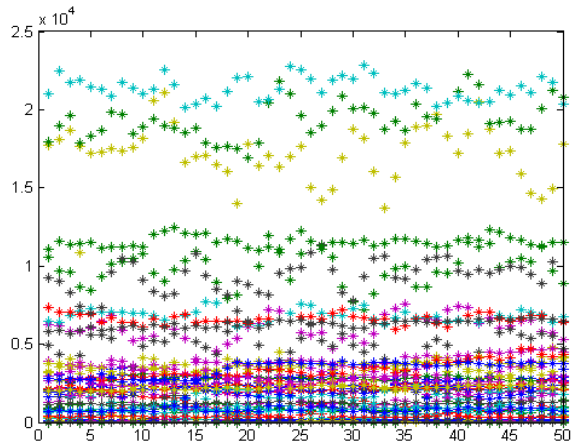


Fig 2: Plot of original OD flow from Abilene data set

We implement our proposed detection method using MATLAB, which is a high-level technical computing language and interactive environment for algorithm development, data visualization and analysis.

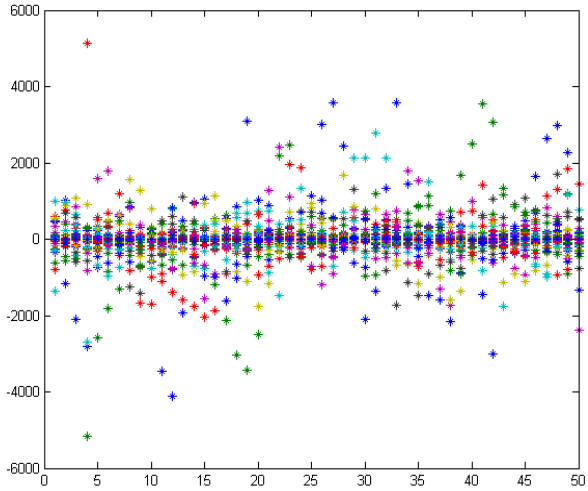


Fig 3: Plot of data after applying our proposed method.

Figure 3 is a plot of Abilene data after applying our proposed method, the normal traffic is centered in the middle, while anomalies deviates from the behavior or normal traffic, it tends to scatter far from the center.

To evaluate our algorithm, we examined its performance on network-wide traffic datasets analyzed by Lakhina et al. in [2], with well known and identified anomalies, thus we have "ground truth" anomaly annotations against which to compare the output of our Combined PCA and Adaptive Filtering Approach. We found that our Combined PCA and Adaptive Filtering method provides high detection rate, with the added advantage of lower complexity. The experimental results show that our Combined PCA and Adaptive Filtering method detects 85% of anomalies in the Abilene data set.

REFERENCES

[1] A. Lakhina, M. Crovella, and C. Diot, "Mining Anomalies Using Traffic

Feature Distributions," in Proc. SIGCOMM, Philadelphia, PA, Aug. 2005.

[2] Lakhina, A., Crovella, M., and Diot, C. Diagnosing network-wide traffic anomalies. In ACM SIGCOMM (Portland, Oregon, USA, 2004), pp. 219–230.

[3] Hotelling, H. Analysis of a complex of statistical variables into principal components. J. Educ. Psy. (1933), 417–441.

[4] Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E. D., and Taft, N. Structural analysis of network traffic flows. In ACM SIGMETRICS (New York, NY, USA, 2004), pp. 61–72.

[5] Lakhina, A., Crovella, M., and Diot, C. Characterization of network wide anomalies in traffic flows. In ACM Internet Measurement Conference (Taormina, Sicily, Italy, 2004), pp. 201–206.

[6] I.T. Jolliffe, "Principal Component Analysis", 2nd Ed., Springer-Verlag, NY, 2002.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", China Machine Press, Beijing, 2nd edition, 2004.

[8] R.Heady,G.Luger,A.Maccabe "The architecture of network level intrusion detection system", Technical report ,Computer Science Department ,University of New Mexico , August 1990.

[9] F. Feather, D. Siewiorek, and R. Maxion. Fault detection in an ethernet network using anomaly signature matching. In Proceedings of ACM SIGCOMM, 1993.

[10] I. Katzela and M. Schwartz. Schemes for fault identification in communication networks. IEEE/ACM Transactions on Networking, 3(6), Dec. 1995.

[11] M. Thottan and C. Ji. Anomaly detection in IP networks. IEEE Transactions in Signal Processing, 51(8), Aug. 2003.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500