Dimensionality reduction using Principal Component Analysis for network intrusion detection

Article *in* Perspectives in Science · July 2016

DOI: 10.1016/j.pisc.2016.05.010

CITATIONS
23

READS
341

2 authors:

Keerthi Vasan K
National Institute of Technology Puducherry
4 PUBLICATIONS **26** CITATIONS

SEE PROFILE

Surendiran Balasubramanian
National Institute of Technology Puducherry
24 PUBLICATIONS **137** CITATIONS

SEE PROFILE

Perspectives in Science (2016) 8, 510—512

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc

Dimensionality reduction using Principal Component Analysis for network intrusion detection☆

K. Keerthi Vasan*, B. Surendiran

Department of Computer Science, National Institute of Technology (NIT) Puducherry, Karaikal, India

Received 19 February 2016; accepted 31 May 2016
Available online 1 July 2016

KEYWORDS
Intrusion detection;
Dimensionality reduction;
Principal Component Analysis;
KDD CUP;
UNB ISCX

Summary Intrusion detection is the identification of malicious activities in a given network by analyzing its traffic. Data mining techniques used for this analysis study the traffic traces and identify hostile flows in the traffic. Dimensionality reduction in data mining focuses on representing data with minimum number of dimensions such that its properties are not lost and hence reducing the underlying complexity in processing the data. Principal Component Analysis (PCA) is one of the prominent dimensionality reduction techniques widely used in network traffic analysis. In this paper, we focus on the efficiency of PCA for intrusion detection and determine its Reduction Ratio (RR), ideal number of Principal Components needed for intrusion detection and the impact of noisy data on PCA. We carried out experiments with PCA using various classifier algorithms on two benchmark datasets namely, KDD CUP and UNB ISCX. Experiments show that the first 10 Principal Components are effective for classification. The classification accuracy for 10 Principal Components is about 99.7% and 98.8%, nearly same as the accuracy obtained using original 41 features for KDD and 28 features for ISCX, respectively.

Introduction

Traffic analysis for a network is performed to understand the nature of its traffic, monitor the activities and manage hostility in traffic flows. Hostile traffic targeted towards a host intends to either flood the target with large number of packets thereby stalling it is functioning, or probe the target to study it is vulnerabilities or a precise attack on the target by exploiting the vulnerabilities. Signature-based intrusion detection is inefficient (Brauckhoff et al., 2009; Issariyapat and Kensuke, 2009; Vasudevan et al., 2011) since the attacks are generated with sophisticated tools and signatures derived for a session becomes obsolete for another session. Data mining is employed for intrusion detection whereby useful patterns in the traffic can be used to for

☆ This article belongs to the special issue on ''Engineering and Material Sciences''.

* Corresponding author. Tel.: +91 9843656836.
E-mail addresses: keerthivasankkv@gmail.com (K. Keerthi Vasan), surendiran@nitpy.ac.in (B. Surendiran).

http://dx.doi.org/10.1016/j.pisc.2016.05.010

further analysis. Machine learning algorithms study these patterns and develop a model for analyzing new traffic samples. The dataset (KDD, 2016; UNB, 2016) generated from the traffic traces consists of d features and when d is large, the complexity involves processing also high. Dimensionality reduction helps to identify k significant features such that $k < d$ and they achieve classification accuracy same as that of d features. Principal Component Analysis (PCA) is a dimensionality reduction technique which has been used prominently in the field of traffic analysis (Zhang et al., 2012). The organization of the paper is as follows, '*Principal Component Analysis*' section discusses about Principal Component Analysis and existing literature regarding PCA, the results of experiments done are in '*Experiments and results*' section and paper concludes by in '*Conclusion*' section.

Principal Component Analysis

Principal Component Analysis is a feature extraction technique that generates new features which are linear combination of the initial features. PCA maps each instance of the given dataset present in a d dimensional space to a k dimensional subspace such that $k < d$. The set of k new dimensions generated are called the Principal Components (PC) and each principal component is directed towards maximum variance excluding the variance already accounted for in all its preceding components. Subsequently, the first component covers the maximum variance and each component that follows it covers lesser value of variance. The Principal Components can be represented as the following

$$PC_i = a_1X_1 + a_2X_2 + \cdots + a_dX_d \quad (1)$$

where PC_i — Principal Component 'i'; X_j — original feature 'j'; a_j — numerical coefficient for X_j.

PCA is one of the most prominently used feature extraction methods for traffic analysis. Brauckhoff et al. (2009) discuss about implementing PCA with KL expansion method for anomaly detection and issue of right number of PC for analysis. Ringberg et al. (2007) discusses the sensitivity of PCA for anomaly detection, issues related to number of PC, impact of anomaly size and gives a comprehensive study of the related issues on Abilene and Geant networks. Issariyapat and Kensuke (2009) discuss about using PCA for MAWI network and using the information from packet header for detecting anomaly.

Experiments and results

The experiments were carried out on labelled datasets generated from real world network traces. The datasets include KDD Cup (KDD, 2016), generated for KDD Cup Contest of 1999 from DARPA 1998 traces and UNB ISCX, from traces generated at University of Brunswick, Canada. The KDD Cup dataset consists of 41 dimensions while the ISCX dataset was extracted using a novel framework (Vasudevan et al., 2011), having 28 dimensions. A subset used consists of 31,279 instances for KDD Cup and 33,746 instances for ISCX. The attack vectors include Flood, Priest and Probe (only KDD). Classifier algorithms used for this work include C4.5 and Random Forest. The code for PCA and the classifiers are available in Weka library.

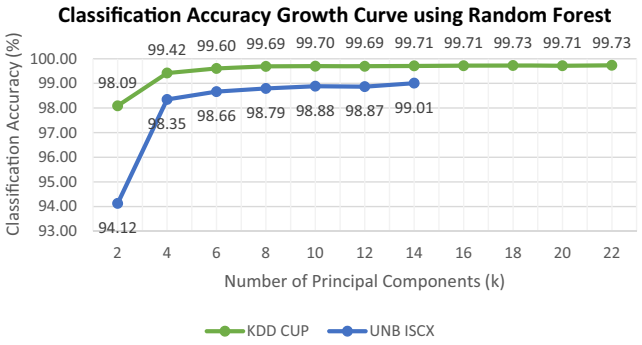

Figure 1 Classification accuracy curve for KDD cup and ISCX dataset with Random Forest.

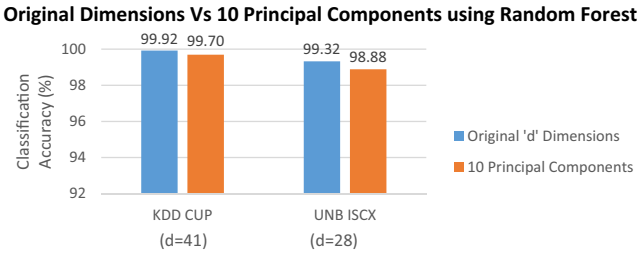

Figure 2 Comparison between original dimensions and 10 Principal Components.

Ideal number of Principal Components

The datasets of size $n \times d$ was mapped to the given k — principal component framework and transformed into dataset of size $n \times k$, where n is number of instances and d is number of original dimensions. Number of Principal Components (PC) represented as k range from 2 to 20, with an incremental offset of two. The Random Forest, a supervised learning algorithm, was applied to each $n \times k$ and a graph was plotted. Fig. 1 depicts the relation between k — PC and classification accuracy. It is observed that initially, the growth of accuracy is more for increase in value of k. After $k = 10$, the growth is either stagnated or insignificant and hence, it is inferred from the plot that the ideal number of PC, $k_{ideal} = 10$.

It is observed from Fig. 2 that accuracy of the classifier, here Random Forest is nearly the same for both k_{ideal} and d dimensions.

Reduction Ratio for Principal Component Analysis

The Reduction Ratio (RR) is the measure for determining the extent of dimensional reduction. RR for PCA (RR_{PCA}) is the ratio of number of target dimensions to number of original dimensions. Lower the value of RR_{PCA}, higher is the efficiency of PCA. RR_{PCA} for our problem of intrusion detection given by the following mathematical expression

$$RR_{PCA} = \frac{k_{ideal}}{d} \quad (2)$$

Given the values of k_{ideal} and d, the RR_{PCA} for KDD Cup dataset is 0.24 and ISCX dataset is 0.36.

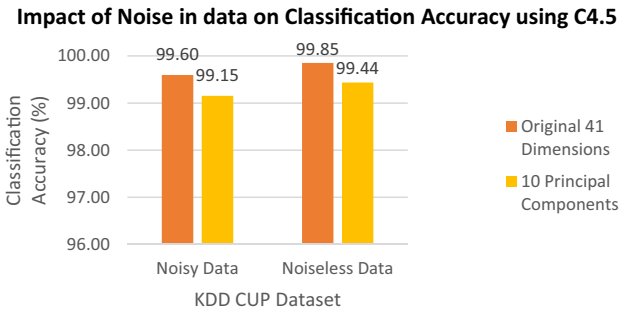

Figure 3 Impact of noise in data using C4.5 classifier.

Impact of noise on performance of PCA

The KDD Cup dataset initially has 39 classes which were mapped to the attack vector mentioned earlier. A new subset consisting of 31,287 instances was used for this analysis. Classes with very few instances were regarded as noise. In Fig. 3, classification accuracy of original d dimensional data and k_{ideal} dimensional data is compared. The difference in classification accuracy of d and k_{ideal} is reduced by 7.3% when noise is eliminated, given C4.5 classifier. Similar reductions of difference in classification were observed for other classifiers also. This shows that PCA improves classification accuracy when data is preprocessed to be noise free.

Conclusion

Principal Component Analysis has shown to be very effective for dimension reduction in intrusion detection. It is identified from experimental results that ideal number of Principal Components (PC) k_{ideal} = 10 for intrusion detection. The classification accuracy of k_{ideal}-PC is nearly equal to that of the original d dimensions. The Reduction Ratio of PCA for KDD Cup and UNB ISCX dataset is 0.24 and 0.36, respectively. Presence of noise degrades classification accuracy and PCA enhances the accuracy when data is noise free. Using PCA for designing an intrusion detection system will reduce the complexity of the system whilst achieving higher classification accuracy.

References

Brauckhoff, D., et al., 2009. Applying PCA for traffic anomaly detection: problems and solutions. In: INFOCOM 2009, IEEE.

Issariyapat, C., Kensuke, F., 2009. Anomaly detection in IP networks with Principal Component Analysis. In: Communications and Information Technology, IEEE.

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (18.02.16).

Ringberg, H., et al., 2007. Sensitivity of PCA for traffic anomaly detection. ACM SIGMETRICS Perform. Eval. Rev. 35 (1).

http://www.unb.ca/research/iscx/dataset/iscx-IDS-dataset.html (18.02.16).

Vasudevan, A.R., et al., 2011. SSENet-2011: a network intrusion detection system dataset and its comparison with KDD CUP 99Dataset. In: II AH-ICI, IEEE.

Zhang, B., et al., 2012. PCA-subspace method — is it good enough for network-wide anomaly detection. In: Network Operations and Management Symposium (NOMS), IEEE.