Improve R2L Attack Detection using Trimmed PCA

ELKHADIR Zyad
LASTID Research Laboratory
Ibn Tofail University, Kenitra
Email: zyad.elkhadir@gmail.com

ARCHI Taha
LASTID Research Laboratory
Ibn Tofail University, Kenitra
Email: architaha1@gmail.com

BENATTOU Mohammed
LASTID Research Laboratory
Ibn Tofail University, Kenitra
Email: mbenattou@yahoo.fr

***Abstract*—Due to the large growth of modern network traffic in term of size and complexity, intrusion detection systems have shown a lot of limits such as the detection rate deterioration and the rising of false positive rate. To overcome this problem, we have to eliminate the noisy content within the original high dimensional data by exploiting a feature extraction method. In literature, many publications proposed to use Principal Component Analysis (PCA). However, this method has an important limitation. The estimated general mean vector is prone to outliers. In this paper, to alleviate this issue, we suggest to exploit the trimmed mean with different percentages. Many experiments on NSL-KDD show a promising results.**

***Keywords*—PCA, Trimmed mean, Network Attack Detection, IDS, NSL-KDD.**

I. Introduction

Network intrusions become a common phenomenon due to the availability and simplicity of automated attack tools. Many threats compromise computer networks and inflict mass level infections and damages such as stealthy crafted zero-day exploits, computer worms and viruses. To handle that, Several preventive mechanisms like encryption, authentication, policy management and firewalls have been proposed in the literature as defensive measures against network intrusions. In addition to these preventive measures, a complementary second line of defense called Intrusion Detection System (IDS) is required to provide a comprehensive security against various network attacks.

Based on their detection technique, IDSs can broadly be classified into following two categories: misuse based and anomaly based.

The first approach uses a set of predefined attack signatures to identify network intrusions. it provides a powerful defense against known attacks, nevertheless it fails in detecting novel attacks. Moreover, the time lapse between the discovery of a new attack and deployment of its corresponding signature is significant. That make misuse based ineffective against some numerous instances of network attacks, notably zero-day exploits and computer worms. In addition, the developed signatures need to be regularly managed, distributed and updated by the security administrator.

The other approach uses a collection of examples that represent normal behavior and constructs a model of familiarity. Therefore, any deviation of the network traffic from the learned model is considered as an anomaly and an alarm is raised whenever any such anomalous network traffic is detected. It main advantage is the ability to detect novel attacks. In this context, various machine learning techniques (MLT) have utilized to build an effective IDS. Examples include Bayesian networks (BN), Markov models, neural networks (NN), fuzzy logic techniques, k-nearest neighbor (k-NN), and support vector machine (SVM). In [1], the authors combined SVM, Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Networks (ANN) to improve the accuracy of the IDS. Similarly, the assembling of radial basis function neural networks (RBF) and decision trees was investigated by [2]. In [3], the authors treated network traffic records as images to ameliorate DoS attack detection. In spite of it all, the anomaly based IDS produces a high false alarm rate since it manipulates large network traffic with useless features.

To tackle with this restriction, many data dimensionality reduction techniques have been exploited. Principal Component Analysis (PCA) is an example of these methods, it projects the original high-dimensional feature space to a low-dimensional space, such that the important features are well preserved. In recent works [4]–[8] , PCA and it variants show promising results. In [4] and [5] KPCA (Kernel Principal Component Analysis) was introduced to extract important features by adopting a non-linear kernel methods, in [6] and [7] the authors used PCA with different norm maximization. In [8] PCA and Fisher Discriminant Ratio (FDR) have been utilized to reduce features and eliminate noise from the network connections. The model was based on probabilistic self-organizing maps (PSOM) and intended to model the feature space and recognize normal from anomalous patterns. The achieved results in term of accuracy, specificity, and sensitivity were promising.

One of the weaknesses of PCA is it sensitivity to outliers, a fact that comes essentially from using arithmetic mean. To address this problem, multiple robust PCA methods have been presented. Among them, there are OMPCA [9], QR-OMPCA [10] and PCA-GM [11], the first one introduces an iterative weighted method which obtain automatically the optimal mean . QR-OMPCA optimizes this process by replacing a SVD decomposition with a more faster and a stable QR-decomposition. PCA-GM is based on the power mean or the generalized mean, which can become the arithmetic, geometric and harmonic means depending on the value of some parameters.

The contribution of this paper is based on using the Trimmed mean vector, to estimate the average vector. This mean vector is more representative of the true central region for skewed data or data with outliers. Thus, the suggested PCA method should be more robust than the previous sample-average based PCA. We will demonstrate this by numerous experiments on NSL-KDD.

The rest of this paper is organized as follows. In Section II,

978-1-5386-8317-0/19/$31.00

we introduce the formulation of PCA. Section III gives details of trimmed PCA. Section IV underlines the description of NSL-KDD, the network simulated database. In Section V we exhibit the experimental results that illustrate the effectiveness of the PCA variant. Finally, Section VI offers our conclusions.

II. Principal Component Analysis

The main purpose of principal component analysis is to decrease the dimensionality of the initial dataset, while retaining as much as possible the variance present in this dataset. This is done by considering only the first few Principal components (PCs), sorted in decreasing order [12].

In mathematical terms, suppose we have a training set of M vectors $w_1, w_2, ., w_M$, each vector contain n features. To get $n'(n' << n)$ principal components of the training set we rely on the following steps:

1) Compute the average σ of this set :

$$\sigma = (\frac{1}{M}) \sum_{i=1}^{M} w_i \tag{1}$$

2) Subtract the mean σ from w_i and get ρ_i :

$$\rho_i = w_i - \sigma \tag{2}$$

3) Compute the covariance matrix C where :

$$C_{n \times n} = (\frac{1}{M}) \sum_{i=1}^{M} \rho_i \rho_i^T = AA^T \tag{3}$$

and

$$A_{n \times M} = (\frac{1}{\sqrt{M}}) \rho_i \tag{4}$$

4) Let U_k be the k^{th} eigenvector of C corresponding to the λ_k associated eigenvalue and $U_{n \times n'} = [U_1 ... U_{n'}]$ the matrix of these eigenvectors, so we have

$$CU_k = \lambda_k U_k \tag{5}$$

5) Sort the eigenvalues (and the corresponding eigenvectors) in decreasing order and choose the first eigenvectors called Principal Components (PCs). Practically, the number of the principal components chosen depends on the precision explicitly expressed by

$$\tau = \frac{\sum_{i=1}^{n'} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{6}$$

This ratio defines the information rate kept from the whole rough input data, by the corresponding n eigenvalues.

From the above formulation of PCA we observe that the mean vector (eq. (1)) has a big role in defining the covariance matrix. Thus, more the mean vector is accurate more the resulting principal components U will be. Nevertheless, there are numerous studies which confirm that sample average may not express the true central region for skewed data or data with outliers.As a consequence, the classical PCA will fail in giving the correct U.

III. Trimmed PCA Formulation

A. Trimmed Mean

Trimmed mean is seen as a statistical measure of central tendency. In order to calculate it, we search the mean of a distribution after removing equal number of samples from the high and the low extremities. The number of samples to be eliminated is mostly given as a percentage P of the total number of samples. Suppose we have M numbers $x_1, x_2, ..., x_M$, here are the steps to find the truncated mean tm:

1) Reorder x_i from the smallest to the largest value.
2) Compute the trimmed proportion $p = P/100$.
3) Compute $k = M \times p$ .If k is a float, round it to the nearest integer.
4) $tm = (\frac{1}{M-2k})(x_{k+1} + x_{k+2} + .. + x_{M-k})$.

tm is less sensitive to outliers than the sample average. Moreover, it is robust and efficient when dealing with mixed distributions and heavy-tailed distribution (like the Cauchy distribution). In case we have a set of vectors or a matrix given by :

$$X = [X_1, X_2, .., X_M] = \begin{bmatrix} X_{11} & X_{21} & X_{31} & \dots & X_{M1} \\ X_{12} & X_{22} & X_{32} & \dots & X_{M2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & X_{3n} & \dots & X_{Mn} \end{bmatrix}$$

It trimmed mean is $tm = (tm_{1n}, tm_{2n}, .., tm_{Mn})$ where tm_{in} is the trimmed mean of $(X_{i1}, X_{i2}, .., X_{in})$ the i-th column of the data matrix X.

B. Trimmed PCA

The proposed PCA variant will consider tm as estimator instead of the mean vector σ and consider the below steps :

1) Compute the trimmed mean tm
2) Subtract tm from w_i' and get ρ_i' :

$$\rho_i' = w_i' - tm \tag{7}$$

3) Compute the covariance matrix C' where :

$$C'_{n \times n} = (\frac{1}{M}) \sum_{i=1}^{M} \rho_i' \rho_i'^T = A' A'^T \tag{8}$$

and

$$A'_{n \times M} = (\frac{1}{\sqrt{M}}) \rho_i' \tag{9}$$

4) Let U_k' be the k^{th} eigenvector of C' corresponding to the λ_k' associated eigenvalue and $U'_{n \times n'} = [U'_1 ... U'_{n'}]$ the matrix of these eigenvectors, so we have

$$C'U_k' = \lambda_k' U_k' \tag{10}$$

In the same way, we rely on τ' to choose the number of the new principal components :

$$\tau' = \frac{\sum_{i=1}^{n'} \lambda_i'}{\sum_{i=1}^{n} \lambda_i'} \tag{11}$$

The projection of a new vector x_{new} on the space constructed by our approach is obtained by:

$$t_i = (U_i')^T x_{new} \tag{12}$$

Hereafter the algorithm is called trimmed PCA.

IV. THE SIMULATED DATABASE

NSL-KDD is a data set [13] proposed to solve some of the shortcomings of the KDD'99 data set discussed in [14] . The new dataset proposes a reasonable number of train records (125973 samples) and test sets (22544 samples). This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable. In addition, there is no redundancy sample present in the dataset and testing set contains some attack which are not present in the training set.

V. EXPERIMENTS AND DISCUSSION

In this section, many experiments were designed to show the superiority of Trimmed PCA over the classical PCA.

To estimate the accuracy of these methods we employ Detection Rate (DR) and False Positive Rate (FPR):

$$DR = \frac{TP}{TP + FN} \times 100 \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \times 100 \tag{14}$$

True positives (TP) means the number of attacks correctly predicted. False negatives (FN) refer to the intrusions incorrectly classified, false positive (FP) are normal instances wrongly classified, and true negatives (TN) are normal instances classified as normal.

In the experiments, we vary the number of training samples and consider the following composition (100 normal data, 100 DOS data, 50 U2R data, 100 R2L data, and 100 PROBE) as testing data. We rise the number of DOS and PROBE attacks on the one hand, on the other hand, we set normal training data at 1000 samples. U2R and R2L samples are fixed at 50 and 100 respectively. DR and FPR took the average of twenty time random selections. We used the K-nearest neighbor as a classifier with the euclidean and cityblock distances. The classification is done in a subspace generated by the first 4 principal components.

Fig. 1 and Fig. 2 show the results of the comparison using the euclidean distance. In trimmed PCA we investigate the effect of p=5%, p=10% and p=15%. We observe that once 2100 samples is exceeded, the proposed approach takes the advantage, particularly for p=15%. Furthermore, the approach gives fewer FPR. These results prove that arithmetic mean becomes more sensitive in presence of big training data containing outliers.

Figs. 3 and 4 exhibit the results we found when we consider cityblock distance, we can see that the proposed approach still give better results.

Fig. 1: DR of Trimmed PCA and PCA for euclidean Distance

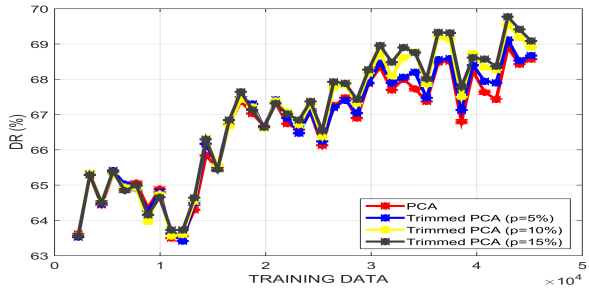


Fig. 2: FPR of Trimmed PCA and PCA for euclidean Distance

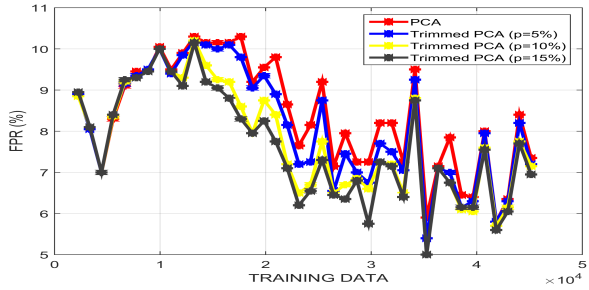
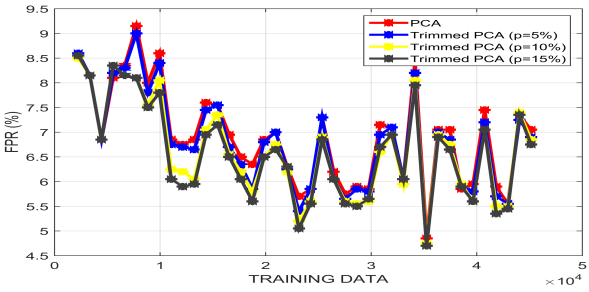


Fig. 3: DR of Trimmed PCA and PCA for cityblock Distance

Fig. 4: FPR of Trimmed PCA and PCA for cityblock Distance



In order to go deeper in our investigation we show in TABLE I and TABLE II the detection rate of every type of attack (DOS,U2R,R2L,PROBE) while considering the euclidean and cityblock distances. We observe that there is at least 1% of difference between the R2L detection rate of PCA and R2L identification rate of Trimmed PCA. It means that even if

TABLE I: Training data size vs. Attack's detection rate (%) for euclidean Distance

	PCA				Trimmed PCA (10%)				Trimmed PCA (15%)			
Training data	DOS	U2R	**R2L**	PROBE	DOS	U2R	**R2L**	PROBE	DOS	U2R	**R2L**	PROBE
31950.0	82.7	13.0	**3.8**	68.6	82.5	13.0	**4.9**	69.2	82.6	13.2	**5.4**	70.1
33050.0	83.7	11.4	**3.1**	63.5	83.7	11.8	**4.8**	64.6	83.6	12.0	**5.3**	65.0
34150.0	83.2	12.2	**3.3**	70.8	83.4	12.2	**5.4**	70.3	83.3	12.2	**6.1**	71.3
35250.0	86.1	14.2	**5.4**	68.8	86.2	14.2	**6.7**	68.6	86.3	14.2	**7.5**	68.1
36350.0	82.7	11.8	**1.2**	66.9	82.7	12.0	**2.8**	67.6	82.7	12.0	**3.1**	67.4
37450.0	83.8	11.8	**3.9**	70.4	83.8	12.0	**5.6**	68.8	83.9	12.0	**6.3**	67.6
38550.0	80.3	12.2	**3.7**	67.9	80.6	12.2	**5.9**	67.8	80.6	12.2	**6.8**	68.2
39650.0	84.0	12.8	**1.4**	68.9	84.1	13.0	**2.9**	69.7	84.2	13.0	**3.6**	70.6
40750.0	83.7	12.6	**1.1**	69.3	83.6	12.6	**2.2**	68.2	83.9	12.8	**2.9**	70.2
41850.0	82.2	10.8	**1.4**	69.7	82.2	10.6	**3.8**	69.0	82.2	10.6	**4.2**	71.0
42950.0	84.0	14.6	**4.7**	68.9	84.1	14.8	**6.3**	70.0	84.1	14.8	**6.9**	71.8
44050.0	83.0	12.6	**3.4**	68.7	83.1	12.6	**4.9**	68.3	83.0	12.6	**5.2**	69.4
45150.0	82.9	10.8	**5.5**	70.3	83.3	10.8	**7.2**	69.7	83.3	10.8	**7.8**	70.7

TABLE II: Training data size vs. Attack's detection rate (%) for cityblock Distance

	PCA				Trimmed PCA (10%)				Trimmed PCA (15%)			
Training data	DOS	U2R	**R2L**	PROBE	DOS	U2R	**R2L**	PROBE	DOS	U2R	**R2L**	PROBE
31950.0	83.9	11.6	**9.7**	66.5	84.0	11.8	**11.9**	69.0	84.1	11.8	**12.0**	68.4
33050.0	83.3	10.4	**7.6**	71.1	83.3	10.6	**9.1**	72.0	83.3	10.8	**9.6**	71.3
34150.0	83.7	11.4	**14.7**	64.8	83.8	11.6	**16.0**	67.7	83.8	11.6	**16.2**	66.6
35250.0	82.5	9.4	**13.3**	71.5	82.7	9.4	**15.3**	71.9	82.8	9.4	**15.5**	72.5
36350.0	84.2	9.8	**11.8**	67.1	84.3	10.0	**13.3**	68.2	84.1	10.0	**12.9**	68.1
37450.0	83.0	9.6	**10.2**	69.4	82.9	10.0	**11.2**	70.1	83.1	10.0	**11.5**	68.8
38550.0	82.9	11.8	**5.5**	67.1	83.0	11.6	**6.8**	67.5	83.0	11.8	**7.9**	66.7
39650.0	84.2	11.6	**9.2**	67.1	84.7	12.6	**10.5**	67.9	84.6	12.6	**10.7**	69.4
40750.0	80.5	9.8	**8.7**	66.6	80.4	10.2	**10.9**	66.8	80.3	10.4	**11.2**	66.4
41850.0	82.9	11.8	**7.5**	71.6	83.0	12.6	**8.3**	70.6	83.0	12.8	**8.6**	69.2
42950.0	82.7	11.2	**8.9**	69.9	82.8	11.4	**10.4**	70.0	82.8	11.4	**10.6**	69.8
44050.0	83.5	9.8	**8.6**	68.7	83.7	10.2	**10.3**	69.1	83.7	10.2	**10.5**	68.1
45150.0	83.0	10.4	**10.2**	68.2	83.1	11.2	**11.8**	69.5	83.1	11.0	**12.1**	70.3

we have a small portion of training data (R2L) in presence of outliers, the proposed approach could achieve promising results.

VI. Conclusion

In this paper, we ameliorate the accuracy of PCA in detecting R2L Attacks, by using class trimmed mean vector, rather than the class sample average. Experiments on NSL-KDD demonstrate the effectiveness of the proposed model, however, the approach has a drawback: The determination of the adequate percentage of data P which has to be discarded is based on an empirical process. To handle that in future works, we will try to automate this process and make it more efficient.

References

[1] S. Mukkamala, A. H. Sung, and A. Abraham, "Intrusion detection using an ensemble of intelligent paradigms," *Journal of network and computer applications*, vol. 28, no. 2, pp. 167–182, 2005.

[2] M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Engineering*, vol. 30, pp. 1–9, 2012.

[3] Z. Tan, A. Jamdagni, X. He, P. Nanda, R. P. Liu, and J. Hu, "Detection of denial-of-service attacks based on computer vision techniques," *IEEE transactions on computers*, 2015.

[4] Z. Elkhadir, K. Chougdali, and M. Benattou, "Intrusion detection system using pca and kernel pca methods," in *Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015*. Springer, 2016, pp. 489–497.

[5] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid kpca and svm with ga model for intrusion detection," *Applied Soft Computing*, vol. 18, pp. 178–184, 2014.

[6] C. Khalid, E. Zyad, and B. Mohammed, "Network intrusion detection system using l1-norm pca," in *Information Assurance and Security (IAS), 2015 11th International Conference on*. IEEE, 2015, pp. 118–122.

[7] Z. Elkhadir, K. Chougdali, and M. Benattou, "Network intrusion detection system using pca by lp-norm maximization based on conjugate gradient," *Int Rev Comput Softw (IRECOS)*, vol. 11, no. 1, pp. 64–71, 2016.

[8] E. De la Hoz, E. De La Hoz, A. Ortiz, J. Ortega, and B. Prieto, "Pca filtering and probabilistic som for network intrusion detection," *Neurocomputing*, vol. 164, pp. 71–81, 2015.

[9] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *International conference on machine learning*, 2014, pp. 1062–1070.

[10] Z. Elkhadir, K. Chougdali, and M. Benattou, "An effective cyber attack detection system based on an improved ompca," in *Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on*. IEEE, 2017, pp. 1–6.

[11] J. Oh and N. Kwak, "Generalized mean for robust principal component analysis," *Pattern Recognition*, vol. 54, pp. 116–127, 2016.

[12] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[13] [Online]. Available: http://nsl.cs.unb.ca/NSL-KDD/

[14] M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*, 2009.