COMPUTERS & SECURITY 86 (2019) 63–74

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

A cyber network attack detection based on GM Median Nearest Neighbors LDA

*Zyad Elkhadir**, *Benattou Mohammed*

Lastid Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

ARTICLE INFO

Article history:
Received 26 February 2019
Revised 25 May 2019
Accepted 27 May 2019
Available online 10 June 2019

Keywords:
Linear discriminant analysis
Median NN-LDA
Generalized mean
Network anomaly detection
Feature extraction methods
NSL-KDD
KDDcup99

ABSTRACT

The continuous development in network technologies causes a considerable hike in number of attacks and intrusions. Identification of these threats has become a critical part of security. To fulfill this task, the Intrusion Detection Systems (IDS) were created. Unfortunately, these tools have curse of dimensionality which tends to increase time complexity and decrease resource utilization. As a consequence, it is desirable that important features of network traffic must be analyzed. To obtain these features, previous work has employed a variant of Linear Discriminant Analysis (LDA) called Median Nearest Neighbors-LDA (Median NN-LDA). This approach finds the relevant features by working with network connections that are near to the median of every class. However, Median NN-LDA has an important drawback. It employs the class arithmetic mean vectors in the within and between scatter matrices formulation. As the arithmetic mean is sensitive to outliers, the approach will not produce optimal results. To deal with that, this paper introduces a new robust Median NN-LDA based on the generalized mean. Many experiments on KDDcup99 and NSL-KDD indicate the superiority of the approach over many LDA variants.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Intrusion can be defined as the entire procedure trying to compromise the confidentiality, integrity, or availability (CIA) of information resources. Therefore, it is primordial to employ different measures to avoid such risks. To do that, Denning (1987) invents the concept of detecting cyber-based attacks on computer networks by constructing a framework for intrusion detection system (IDS), which detect security violations by monitoring system audit records for abnormal patterns of system usage.

Intrusion Detection Systems (IDS) are most of time exploited to capture network packets in order to provide a better explanation of what is happening in a particular network. Two mainstream preferences for IDSs are (1) host-based IDSs, and (2) network-based IDSs. The detection methods used in IDS are anomaly based and misuse based (also called signature or knowledge based). Every IDS type has it own advantages and restrictions. In misuse-based detection, data gathered from the system is compared to a set of rules or patterns, also known as signatures. If the data corresponds to one or many signatures, the IDS produce alarms. For detecting known intrusions misuse-based seems to be very powerful, nonetheless, it does not contribute much in terms of zero day attack detection. Anomaly-based IDS on the other hand exploits collections of data containing examples of normal behavior and builds a model of familiarity. Therefore, any action that deviates from the model is considered suspicious and is considered as an intrusion. This type of approaches manages well the zero day attacks. Nevertheless, this approach may produce an important number of false alarms, furthermore, it detection rate is quite satisfactory. That happens due to

* Corresponding author.
E-mail address: zyad.elkhadir@gmail.com (Z. Elkhadir).
https://doi.org/10.1016/j.cose.2019.05.021
0167-4048/© 2019 Elsevier Ltd. All rights reserved.

manipulating enormous network traffic containing many irrelevant features.

To deal with the high dimension data issue in general, many useful dimension reduction techniques have been developed, such as principal component analysis (PCA) (Jolliffe, 2002), linear discriminant analysis (LDA) (Duda et al., 2012), maximum margin criterion (MMC) (Li et al., 2004), regularized discriminant analysis (RDA) (Dai and Yuen, 2007), locality preserving projections (LPP) (He et al., 2005), and marginal fisher analysis (MFA) (Yan et al., 2007). Recently, researchers introduced many ameliorated dimension reduction variants. Lu and Tan (2010) proposed a parametric regularized LPP, Deng et al. proposed a transform-invariant PCA (Deng et al., 2014). This approach was applied on face recognition, it characterizes precisely the inherent structures of the human face which are insensitive to the in-plane transformations. Lu et al. proposed a PCA method (Lu et al., 2016) which achieves dimension reduction and variable selection in same time. This technique gives a better interpretation of the obtained results. Several studies utilized L1-norm instead of L2-norm in the formulation of optimization problem to improve the robustness of PCA against outliers (Ding et al., 2006; Ke and Kanade, 2005; Kwak, 2008). In Ke and Kanade (2005), L1-norm was used to construct an optimal cost function, meanwhile, a convex programming was proposed. R1-PCA (Ding et al., 2006) was introduced to get a solution characterized by the rotational invariance, which is a primordial property for learning algorithms. In Kwak (2008), PCA-L1 was proposed. This variant obtains the reduced space by maximizing an L1 dispersion. Latter, a general Lp-norm PCA with arbitrary p was proposed in Kwak (2014), Liang et al. (2013), Elkhadir et al. (2016). MaxEnt-PCA (He et al., 2010) maximizes Renyi's quadratic entropy to finds the reduced subspace. To estimate the Renyi's entropy, the authors employed a non-parametric Parzen window technique. In He et al. (2011), the authors developed HQ-PCA which is based on maximum correntropy criterion to obtain the inherent features.

LDA in it side employs the well-known Fisher criterion to extract a linearly independent discriminant vectors and exploit them as basis by which samples are projected into a new space. These vectors contribute in maximizing the ratio of the inter-class distance to intra-class distance in the obtained space. Recent works improved this reduction algorithm. Abou-Moustafa et al. (2015) presented a Pareto LDA in order to maximize the distance which exist between all class means. Ghassabeh et al. proposed a fast incremental LDA Ghassabeh et al. (2015). The latter increases the convergence rate of the algorithm. Wang et al. proposed a semisupervised LDA (Wang et al., 2016), which can exploit a small number of labeled data and a limited unlabeled ones for training. A relevance MFA (Ji et al., 2016) was proposed by Ji et al. to formulate the pairwise constraints of relevance-link and irrelevance-link into the relevance graph and irrelevance graph. Ren et al. proposed the outlier Suppressing LDA (Ren et al., 2015) to explore the necessity of the sample itself in constructing the optimal subspace.

Recent papers in networking literature such in Zhang et al. (2015), Zyad et al. (2017a), Elkhadir et al. (2017), Zyad et al. (2017b) have applied different variants of LDA and show a promising initial results. Zhang et al. (2015) proposed an improved K-means clustering algorithm based on linear discriminant analysis (LDA), called LKM algorithm. The latter applies LDA to divide the high-dimension network traffic into 2-dimension data set, after that, it employs K-means algorithm for clustering analysis of the dimension-reduced data. Experiments show that LKM algorithm decreases the sample feature extraction time and improves the accuracy of K-means clustering algorithm. The publication (Zyad et al., 2017a) consists of combining R1-PCA and median LDA to benefit from many advantages such as more resistance against outliers, providing supplementary data variance within and between classes, giving high discrimination power to the R1-PCA principal components, finally solving the Small Sample Size (SSS) problem encountered by median LDA. Zyad et al. (2017b), improve the accuracy of LDA in detecting cyber Attacks, by using class truncated mean vector, rather than the class sample average in the between and within class matrices. The same authors introduced a novel LDA variant called Median NN-LDA (Elkhadir et al., 2017). In this LDA approach they exploit the median of every class to compute the within and between scatter matrices. As a consequence, the approach preserves the local and the global distributions in one hand. In the other hand, it provides more resistance against outliers. Therefore, the proposed method is more robust than traditional linear discriminant analysis methods. To show the effectiveness of these approaches the authors conduct a set of experiments on KDDcup99 and NSL-KDD.

Median NN-LDA has an important drawback. It employs the class arithmetic mean vectors in it mathematical formulation. As the arithmetic mean is sensitive to outliers, the approach will not produce optimal results. To deal with that, inspired by Oh et al. (2013) and Oh and Kwak (2016) this paper introduces a new robust Median NN-LDA based on the generalized mean (Bullen, 2003). The proposed method, General Median Nearest Neighbor LDA (GM Median NN LDA), is a generalization of Median NN-LDA by replacing the arithmetic means with the generalized means. The proposed method can effectively prevent outliers from dominating Fisher criterion by controlling the parameter in the generalized mean. Moreover, it is rotational invariant as it still uses the Euclidean distance as the distance measure in between an within scatter matrices.

The rest of this paper is organized as follows. The next section provides a brief overview of LDA and Median NN-LDA. Section 3 presents in details the proposed approach. Section 4 introduces the two well known network datasets KDDcup99 and NSL-KDD. Section 5 illustrates the IDS architecture on which this research is based. In Section 6 we give the experimental results and illustrate the effectiveness of the algorithm and compare it to some of LDA variants. Finally, Section 7 offers our conclusions.

2. Related work

2.1. *Linear discriminant analysis (LDA)*

LDA tries to reduce dimensionality while keeping the maximum of class-discriminatory information. Its projects original data onto a lower dimensional space which respects maximizing separation of different classes and minimizing dispersion

of samples of the same class simultaneously. In mathematics terms, suppose we have a data matrix $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ composed of n samples, we would find a linear transformation $G \in \mathbb{R}^{d\times l}$ that transforms each vector x_i to a new vector x_i^l in the reduced l-dimensional space as follows:

$$x_i^l = G^T x_i \in \mathbb{R}^l (l < d)$$

The data matrix X can be expressed as $X = [X_1, \ldots, X_k]$ such that k is the number of classes and $X_i \in \mathbb{R}^{d\times n_i}$ represents samples of the ith class, n_i is the sample size of the ith class and $\sum_{i=1}^{k} n_i = n$. LDA operates on two important matrices namely within-class, between-class which are defined as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in X_i}(x - c_i)(x - c_i)^T \quad (1)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - c)(c_i - c)^T \quad (2)$$

c_i is the mean of the ith class, and c is the general mean. From (1) and (2) we obtain:

$$trace(S_w) = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in X_i} ||x - c_i||^2 \quad (3)$$

$$trace(S_b) = \frac{1}{n}\sum_{i=1}^{k} n_i ||c_i - c||^2 \quad (4)$$

The trace of S_w expresses the closeness of every sample to it class mean. The trace of S_b shows us how each class is far from the global mean. In the dimensionality reduced space transformed by G, the two scatter matrices become:

$$S_w^l = G^T S_w G$$
$$S_b^l = G^T S_b G$$

The optimal projection matrix can be gained by maximizing the following objective function or the Fisher criterion given by:

$$G = \arg\max \frac{trace(S_b)}{trace(S_w)} \quad (5)$$

When S_w is invertible, the solutions to (5) can be obtained by performing the following generalized eigenvalue decomposition:

$$S_w^{-1} S_b g_i = \lambda_i g_i \quad (6)$$

Where $G = [g_1, \ldots, g_l]$.

LDA uses the global structure information of the total training samples to determine the linear discriminant vectors. That leads to erroneous results as the global distribution of the data does not represent the real distribution nature of every class. Furthermore, the method is sensitive to outliers.

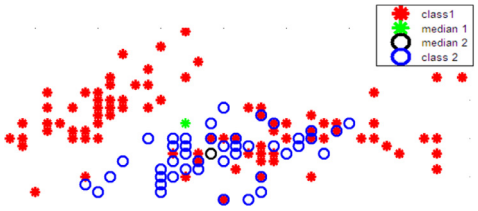


Fig. 1 – Two non linear classes.

2.2. *Median nearest neighbors LDA (Median NN-LDA)*

Inconsistent classes are classes whose true structures can not be defined. They may be non-Gaussian or non-linearly separable. This type of class causes a problem to LDA. To overcome this limit, there is a tendency to divide the class into two distributions. A central or local distribution that defines in a certain way the nature of the class and a global distribution that determines the boundaries of the class. These 2 distributions must be preserved when obtaining the projection vectors. In the literature, (LFDA) has been proposed (Sugiyama, 2007) then its semi-supervised version (Sugiyama et al., 2010) was introduced in 2010. These methods consider local inter-class and intra-class matrices. To calculate them, the first approach uses the weighting matrices $W(lb)$ and $W(lw)$. The semi-supervised version calculates the matrix of regularized weighting $W(rlb)$. These matrices are based on a factor b which determines the elements belonging to the same class and those that do not belong. The authors claim that the structure of local data in the same class tends to be preserved when b is small, but it is not more preserved when b is big. However, to calculate the matrices of weights, we necessitate a lot of time. The work (Chen et al., 2005) was based on the preservation of the neighborhood relationship during the projection of the classes in a space of reduced dimension. The neighboring points represent the local distribution while the others constitute the global distribution. The authors start by building neighborhood graphs G and $G(0)$, then they calculate the affinity weights to define the degree of connectivity of each class element. Finally, they perform a eigenvalue decomposition to find the projection matrix. LPMIP (Wang et al., 2008) seeks a compromise between global and local structures, which is adjusted by a parameter a. Unfortunately, it follows the same philosophy as that used in previous works. It has to do a large calculation of the weighting matrices.

Among the methods which solve the inconsistency problem, there is Median NN-LDA. The next two subsection present the core idea behind the approach.

2.3. *The basic idea*

Fig 1 illustrates two non-consistent and non-Gaussian classes. One in blue and the other in red. To overcome the problem of inconsistency and find local boundaries between these two classes, Median NN-LDA (Elkhadir et al., 2017) proposes to exploit the local distribution of each class. To do that, it relies on the concept of the median (Leys et al., 2013). In the theory of probabilities and statistics, the median is defined as a vector that separates the upper half of a probability

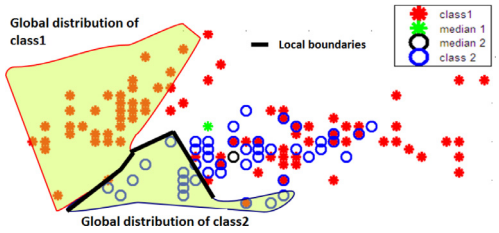


Fig. 2 – The basic idea behind median NN-LDA.

distribution from the lower half. Consequently, the samples which are close to the median represent the central distribution of every class and match logically with the local distribution. In the other hand the further samples represent the global distribution, since they exist naturally in the boundaries of the class and facilitate the separation of classes. With this concept, Median NN-LDA dissociates the two distributions. Therefore, it resolves the matter of distribution's consistency.

The approach performs perfectly even if the data is not Gaussian or have non-linear boundaries. Since it can extract global structures from data by determining the samples that are far from the median. The process can obtain a certain number of local linear discriminant vectors approaching the nonlinear boundary between the classes. This process is illustrated in Fig 2.

2.4. *Mathematical formulation*

In mathematical terms, X_i will be divided into X_i^w and X_i^b.

Let $X_i^w = [x_1, \ldots, x_t] \in \mathbb{R}^{d\times t}$ represents the t median nearest neighbors of every class.

Let $X_i^b = [x_{t+1}, \ldots, x_{n_i}] \in \mathbb{R}^{d\times(n_i-t)}$ contains the $n_i - t$ samples which are far from the median of every class.

The local distribution X_i^w will be exploited by the new within class scatter matrix S'_w, since it measures the intra-class compactness. In the other hand, the global distribution represented by X_i^b is required to compute the new between-class scatter matrix S'_b and more specifically the general mean c.

Then the Eqs. (1) and (2) will be rewritten as follow:

$$S'_w = \frac{1}{t}\sum_{i=1}^{k}\sum_{x\in X_i^w}(x - c_i^w)(x - c_i^w)^T \tag{7}$$

$$S'_b = \frac{1}{t}\sum_{i=1}^{k}(c_i^w - c)(c_i^w - c)^T \tag{8}$$

Where:

$$c_i^w = \frac{1}{t}\sum_{i=1}^{t}(x_i^w) \tag{9}$$

$$c_i^b = \frac{1}{n_i - t}\sum_{i=1}^{n_i-t}(x_i^b) \tag{10}$$

$$c = \frac{1}{k}\sum_{i=1}^{k}(c_i^b) \tag{11}$$

As a consequence, Eqs. (3) and (4) will be replaced by:

$$trace(S'_w) = \frac{1}{t}\sum_{i=1}^{k}\sum_{x\in X_i^w}||x - c_i^w||^2 \tag{12}$$

$$trace(S'_b) = \frac{1}{t}\sum_{i=1}^{k} n_i||c_i^w - c||^2 \tag{13}$$

Median NN-LDA obtains the discriminant vectors by maximizing the following objective function:

$$G' = \arg\max\frac{trace(S'_b)}{trace(S'_w)} \tag{14}$$

The solution can be reached by performing:

$$(S'_w)^{-1}S'_b g'_i = \lambda'_i g'_i \tag{15}$$

Where $G' = [g'_1, \ldots, g'_l]$.

In order to deal with the singularity problem, the method proposes to apply an intermediate dimensionality reduction stage, such as principal component analysis (PCA) (Jolliffe, 2002) to reduce the data dimensionality before applying median NN-LDA.

3. The proposed method

Median NN-LDA has an important weakness. It exploits the class arithmetic mean vectors in the within and between scatter matrices formulation. As the arithmetic means are prone to outliers, the approach fails in giving optimal projection vectors. To deal with that, inspired by Oh et al. (2013) and Oh and Kwak (2016) this section introduces a new robust Median NN-LDA based on the generalized mean namely GM Median NN-LDA.

3.1. *The generalized mean*

For a $p \neq 0$, the generalized mean m_g of $x_i > 0, i = 1, \ldots, N$ (Bullen, 2003) is defined as:

$$m_g = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^p\right)^{\frac{1}{p}}. \tag{16}$$

The arithmetic mean, the geometric mean, and the harmonic mean are particular cases of the generalized mean when $p = 1$, $p = 0$, and $p = 1$, respectively. In addition, the maximum and the minimum values of the numbers can also be approximated from the generalized mean by making $p \to +\infty$ and $p \to \infty$, respectively. Note that as p decreases (increases), the generalized mean is more affected by the smaller (larger) numbers than the larger (smaller) ones, controlling p gives the ability of modifying the contribution of each number to the generalized mean. This property allows us to handle the existing outliers.

In Oh et al. (2013), it was demonstrated that the generalized mean of a set of positive numbers can be spanned by a linear combination of the elements as follow:

$$m_g = \left(\sum_{i=1}^{N} v_i x_i\right). \tag{17}$$

To find v_i we differentiate this equation with respect to v_i.

$$v_i = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^p\right)^{\frac{1}{p}-1} \times \frac{x_i^{p-1}}{N}. \tag{18}$$

In this article we consider the following equation:

$$\sum_{i=1}^{N} x_i^p = \sum_{i=1}^{N} b_i x_i. \tag{19}$$

Where $b_i = x_i^{p-1}$.

The generalized mean becomes the arithmetic mean when each weight b_i has the same value of 1 if $p = 1$. When $p < 1$, the weight b_i increases as x_i decreases. As a result, the generalized mean is more influenced by the small numbers in x_i in this case. Furthermore, the extent of the influence increases when p decreases.

3.2. *Generalized mean median NN-LDA (GM Median NN-LDA)*

The conventional sample means in Eqs. (9) and (10) can be considered in the sense of the least square as follow:

$$c_i^w = \arg\min \frac{1}{t}\sum_{i=1}^{t} ||x_i^w - c_i^w||_2^2 \tag{20}$$

$$c_i^b = \arg\min \frac{1}{n_i - t}\sum_{i=1}^{n_i - t} ||x_i^b - c_i^b||_2^2 \tag{21}$$

In (20), (21) a small number of outliers in the training samples dominate the objective functions as they are based on the squared distances. Consequently, median NN-LDA leads to a biased projection matrix.

To obtain a robust sample means in the presence of outliers, a new optimization problem is formulated by replacing the arithmetic means in (20) and (21) with the generalized means m_i^w and m_i^w defined as:

$$m_i^w = \arg\min \left(\frac{1}{t}\sum_{i=1}^{t}(||x_i^w - m_i^w||_2^2)^p\right)^{1/p} \tag{22}$$

$$m_i^b = \arg\min \left(\frac{1}{n_i - t}\sum_{i=1}^{n_i - t}(||x_i^b - m_i^b||_2^2)^p\right)^{1/p} \tag{23}$$

These modifications will directly affect Eq. (11). It will be rewritten as:

$$m_g = \frac{1}{k}\sum_{i=1}^{k}(m_i^b) \tag{24}$$

As a consequence, Eqs. (7) and (8) will be replaced by:

$$S_w^g = \frac{1}{t}\sum_{i=1}^{k}\sum_{x \in X_i^w}(x - m_i^w)(x - m_i^w)^T \tag{25}$$

$$S_b^g = \frac{1}{t}\sum_{i=1}^{k}(m_i^b - m_g)(m_i^b - m_g)^T \tag{26}$$

To obtain the projection matrix, the proposed LDA variant maximizes the new Fischer criterion given by:

$$G' = \arg\max \frac{trace(S_b^g)}{trace(S_w^g)} \tag{27}$$

First of all, we should solve the Eqs. (22) and (23). These problems are equivalent to (20) and (21) if $p = 1$. The negative effect of outliers can be alleviated if $p < 1$. Using the fact that x_i^p with $p > 0$ is a monotonic increasing function of x_i for $x_i > 0$, these problems can be converted to

$$m_i^w = \arg\min \sum_{i=1}^{t}(||x_i^w - m_i^w||_2^2)^p \tag{28}$$

$$m_i^b = \arg\min \sum_{i=1}^{n_i - t}(||x_i^b - m_i^b||_2^2)^p \tag{29}$$

In this paper, only positive values of p are taken into consideration. In order to solve these two problems it is sufficient to solve the general form:

$$m = \arg\min \sum_{i=1}^{n}(||x_i - m||_2^2)^p \tag{30}$$

In order to obtain the solution, the following derivation is performed:

$$\frac{\partial}{\partial m}\sum_{i=1}^{n}(||x_i - m||_2^2)^p = 0 \tag{31}$$

Nevertheless, it is difficult to obtain a closed-form solution of the above equation. Using a gradient-based iterative algorithm in this case could have a slow convergence speed. As an alternative, we introduce a new method based on (19) and similar to Oh and Kwak (2016). The latter solves (30) in an iterative way.

In the derivation, we decompose (30) into the form of (19) and consider the weight b_i in (19) as a constant. Then, (30) can be approximated by a quadratic function of $||x_i - m||_2$ which can easily be optimized. The details are as follows. Let us denote the value of m after *iter* iterations as $m^{(iter)}$ The first step of the update rule is, for m close to a fixed $m^{(iter)}$, to represent the objective function in (30) as a linear combination of $||x_i - m^{(iter)}||_2^2$ using (19).

$$\sum_{i=1}^{n}(||x_i - m||_2^2)^p \approx \sum_{i=1}^{n}\alpha_i^{(iter)}||x_i - m||_2^2$$

Where

$$\alpha_i^{(iter)} = (||x_i - m^{(iter)}||_2^2)^{p-1}. \tag{32}$$

Here, the approximation becomes exact when $m = m^{(iter)}$. Note that the objective function near $m^{(iter)}$ can be approximated as a quadratic function of m without computing the Hessian matrix of the objective function. The next step is to find $m^{(iter+1)}$ that minimizes the approximated function based on the computed $\alpha_i^{(iter)}$.

$$\frac{\partial}{\partial m}\sum_{i=1}^{n}\alpha_i^{(iter)}||x_i - m||_2^2 = 0$$

The solution of this equation is just the weighted average of the samples as follows:

$$m^{(iter+1)} = \frac{1}{\sum_{j=1}^{n}\alpha_j^{(iter)}}\sum_{i=1}^{n}\alpha_i^{(iter)}x_i \tag{33}$$

To demonstrate the robustness of the generalized sample mean obtained by Algorithm 1, the experiment consists

Algorithm 1 Generalized sample mean.

1. **Input:** data matrix X and p
2. $iter \leftarrow 0$.
3. $m^{(iter)} \leftarrow m$.
4. **repeat**
5. **Approximation:** for fixed $m^{(iter)}$, compute $\alpha_1^{(iter)}, ..., \alpha_n^{(iter)}$ according to (32).
6. **Minimization:** Using the computed $\alpha_1^{(iter)}, ..., \alpha_n^{(iter)}$ update $m^{(iter+1)}$ according to (33).
7. $iter \leftarrow iter + 1$.
8. **Until:** A stop criterion is satisfied.
9. **Output** $m = m^{(iter)}$

in randomly generating 50 samples from a three-dimensional Gaussian distribution with the mean $m_i = [0, 0, 0]$ and covariance matrix $\sigma_i = [0.5\ 0\ 0;0\ 0.5\ 0;0\ 0\ 0.5]$ for inliers and also generated 10 samples from another three-dimensional Gaussian distribution with the mean $m_o = [7,7,7]$ and covariance matrix $\sigma_o = [0.3\ 0\ 0;0\ 0.3\ 0;0\ 0\ 0.3]$ for outliers. Using the generated samples, the sample mean was computed and two generalized sample means were also obtained by Algorithm 1 with $p = 0.3$ and $p = 0.7$, respectively. Fig. 3 shows the arithmetic sample mean and the two generalized sample means together with the generated samples. We observe clearly that the generalized sample means are placed close to the mean of the inliers [0,0,0] in one hand. In the other hand, the arithmetic sample mean is much more affected by the ten outliers. This fact shows that the generalized sample mean with an adequate value of p is more resistant to outliers than the arithmetic sample mean.

4. The simulated databases

4.1. *KDDcup99*

The KDDcup99 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html intrusion detection datasets relies on the 1998 DARPA initiative, which offers to researchers in intrusion detection field a benchmark where to evaluate various approaches. This dataset is composed of many connections.

A connection is a sequence of TCP packets which begins and ends at some well defined times. In this laps of time, a data flows from a source IP address to a target IP address under a defined protocol.

Every connection is composed of 41 features and it is labeled as normal or malicious. if the connection is malicious, it falls into one of four categories:

1. Probing: surveillance and other probing, e.g., port scanning;
2. U2R: unauthorized access to local superuser (root) privileges, e.g., various buffer overflow attacks;
3. DOS: denial-of-service, e.g. syn flooding;
4. R2L: unauthorized access from a remote machine, e.g. password guessing.

We have worked with "kddcup.data_10_percent" as training dataset and "corrected" as testing dataset. The training set contains 494,021 records which is divided as follow: 97,280 are normal connection records, the rest corresponds to attacks. In the other side, the test set contains 311,029 records composed of 60,593 normal connections. It is important to note that:

1. the test data probability distribution is not like that of the training data;
2. the test data contains some new kind of attacks which are dispersed as follow: 4 U2R attack types, 4 DOS attack, 7 R2L attack and 2 Probing attacks. All these attacks do not belong to the training dataset, a fact that makes the IDS's work more challenging.

4.2. *NSL-KDD*

NSL-KDD https://www.unb.ca/cic/datasets/nsl.html is a new version of KDDcup99 dataset. This dataset has some advantages over the old one and has addressed some of it critical problems. here are the important ones:

1. Duplicate records from the training set are removed.
2. Redundant records from the test set are eliminated to improve the intrusion detection performance.
3. Each difficulty level group contains a number of records which is inversely proportional to the percentage of records in the original KDD data set. As a consequence, we will have a more precise evaluation of different machine learning techniques.
4. It is possible to exploit the complete dataset without selecting a random small portion of data because the number of records in the train and test sets are acceptable. Consequently, evaluation results of different research works will be consistent and comparable.

5. The IDS architecture

We assert that the proposed IDS model will be efficient against every TCP/IP structured cyber network attack which can be converted to numerical features. So after that, we can extract

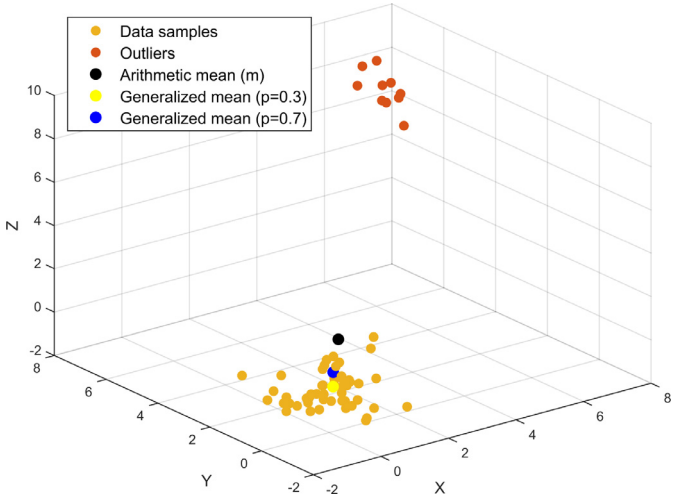


Fig. 3 – 3D toy example with 50 inliers and 10 outliers. The arithmetic mean (m) and the generalized sample means are marked.

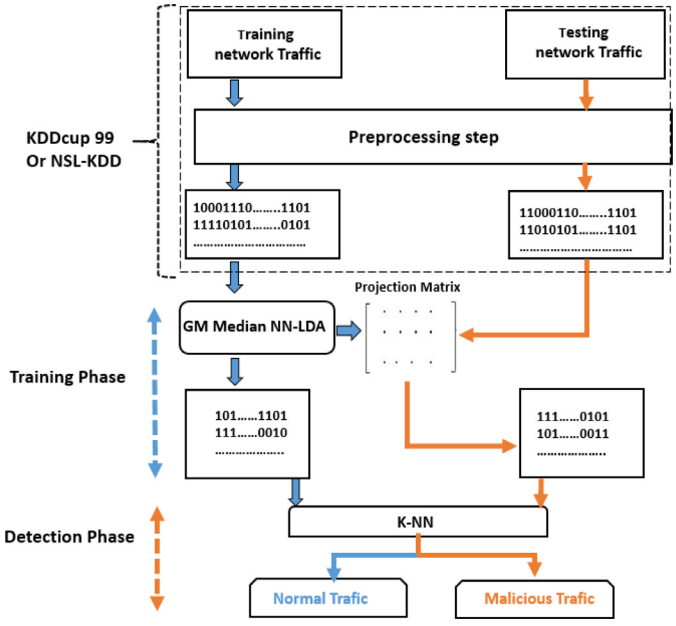


Fig. 4 – The proposed IDS model.

just the valuable features contained in cyber attacks. These features may or may not be directly related to the actual used metrics or attributes such as CPU consumed time, number of web pages visited during a session, the used protocol or service. However, these features will contribute in reducing the different representation spaces before applying some machine learning algorithms such Support Vector Machines, Artificial Neural Networks,..etc. Moreover, they permit to decrease the detection CPU time while keeping or improving the IDS precision.

The IDS architecture is illustrated in Fig. 4. First of all, training data and testing data are taken randomly from KDDcup99 or NSL-KDD (a TCP/IP network connections).

These two parts of raw dataset are preprocessed in order to have a standard feature format. As known, the datasets are defined by continuous and discrete attributes values. The latter have been transformed to continuous values by applying the following transformation. If a discrete attribute i has k values. we correspond i to k coordinates composed of one's and zero's. After that, one coordinate will correspond to every possible value of the attribute. If we consider the flag type attribute which can take the following discrete attributes OTH, REJ, RSTO, RSTO0, RSTR, S0, S1, S2, S3, SF, SH. According to the idea, there will be 11 coordinates for this attribute. As a consequence, suppose a connection record contains a OTH (resp. REJ or RSTO...etc) then the corresponding coordinates will be (1,0,0,0,0,0,0,0,0,0,0) (resp. (0,1,0,0,0,0,0,0,0,0,0) or (0,0,1,0,0,0,0,0,0,0,0)...etc). With this technique, each connection record in the dataset will be represented by 125 coordinates.

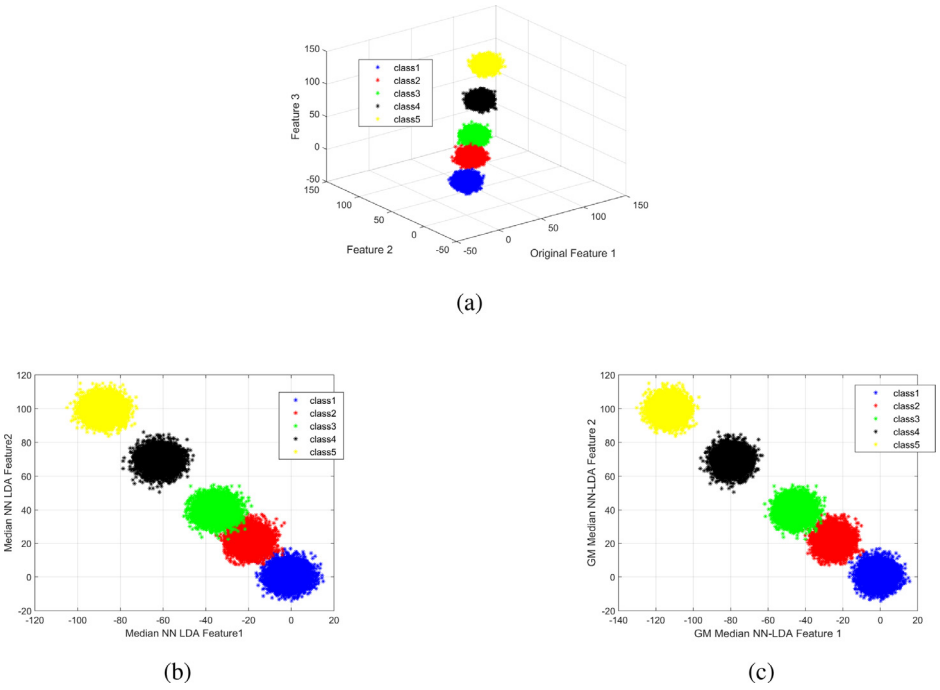


Fig. 5 – A synthetic example. (a) Original 3D synthetic data. (b) Median NN-LDA projection. (c) GM Median NN-LDA projection.

In the training phase, GM Median NN-LDA is used to extract the optimal features from the original high dimensional data, get a new training subset and a projection matrix G'. During the testing phase, G'. will be required to obtain a new testing subset.

Finally, since the goal is evaluating the efficacy of feature extraction method, we use a simple classifier, the K-nearest neighbor classifier KNN (Cover and Hart, 1967) to decide whether the testing samples are normal or not.

6. Experiments and discussion

In this section, first of all, many experiments were conducted on a synthetic data to show the efficiency of GM Median NN-LDA compared to median NN-LDA. After that, a serie of simulations were performed on the two well known benchmark network databases KDDcup99 and NSL-KDD, to demonstrate the performance of the proposed approach method, and compare it with several traditional LDA algorithms, such as LDA (Duda et al., 2012), median LDA (Yang et al., 2006), median NN-LDA (Elkhadir et al., 2017), truncated mean LDA (Zyad et al., 2017b), L1-LDA (Oh and Kwak, 2013), R1-PCA+median LDA (Zyad et al., 2017a).

6.1. *Experiment on a simple instance*

In this subsection, a numerical instance using synthetic data of three dimensions to visually demonstrate the efficiency of GM Median NN-LDA over Median NN-LDA is presented. The synthetic 3D data is composed of multivariate Gaussian five classes. All the classes contains 7000 samples. It should be noted that the class mean and covariance matrices as m_i and σ_i. We adopt the following settings to generate the data. All σ_i equals [20 0 0;0 20 0;0 0 20]. m_1 = [7,7,7], m_2 = [22,22,22], m_3 = [40,40,40], m_4 = [70,70,70] and m_5 = [100,100,100]. The synthetic 3D data are shown as Fig. 5(a)

The synthetic 3D data is projected into the 2D subspace of Median NN-LDA and GM Median NN-LDA. Fig. 5(b) and (c) show the effects after projection. As can be seen, the classes after GM Median NN-LDA projection are much separated and less scattered than these given by Median NN-LDA. We observe that the proposed approach produces a bigger gap between class1 and class2, minimizes the overlap between class2 and class3 in one hand and maximizes the separation between class3, class4 and class5 in the other hand.

6.2. *The KDDcup99 and NSL-KDD results*

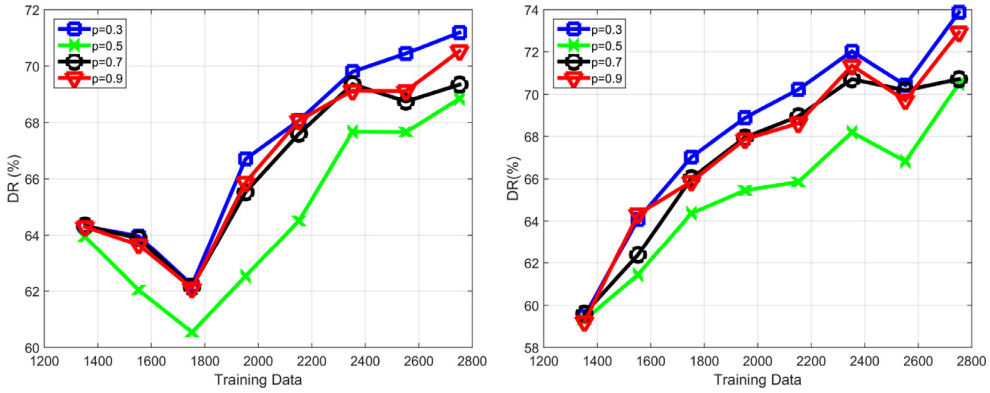
In the experiments, the size of training samples is varying while keeping test dataset intact with the following composition (100 normal data, 100 DOS data, 50 U2R data, 100 R2L data, and 100 PROBE).

The following measures are used to evaluate the LDA variants:

$$DR = \frac{TP}{TP + FN} \times 100 \tag{34}$$

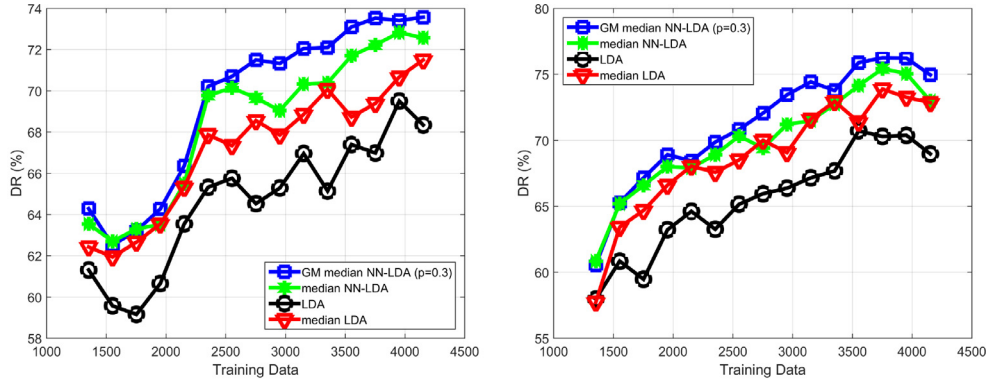
$$FPR = \frac{FP}{FP + TN} \times 100 \tag{35}$$

In network security field, (DR) refers to Detection Rate and (FPR) is False Positive Rate. True positives (TP) are attacks correctly predicted. False negatives (FN) represent intrusions classified as normal instances, false positive (FP) refer to normal instances wrongly classified, and true negatives (TN) are

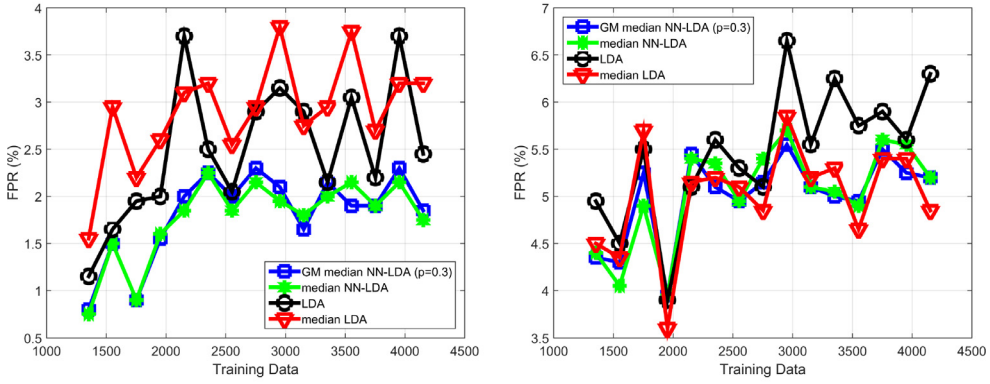


(a) : Results on KDDcup99 (b) : Results on NSL-KDD

Fig. 6 – Evaluation of different values of *p*.



(a) : DR vs Training data using KDDcup99 (b) : DR vs Training data using NSL-KDD



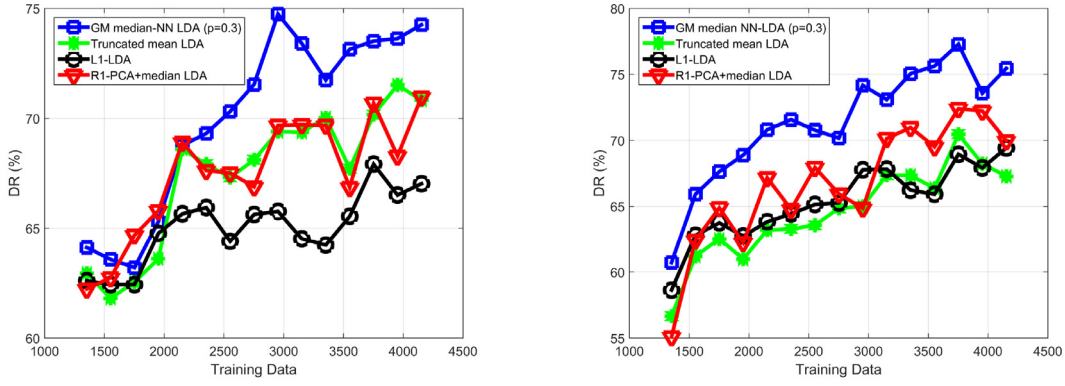
(c) : FPR vs Training data using KDDcup99 (d) : FPR vs Training data using NSL-KDD

Fig. 7 – Comparison of GM median NN LDA with median NN-LDA, LDA and median LDA in term of DR and FPR.

normal instances classified as normal. Therefore, the most reliable feature extraction method, is the one which produces a high DR and a low FPR. To reduce the variation of the detection rate (DR) and (FPR), the mean of twenty runs is adopted.

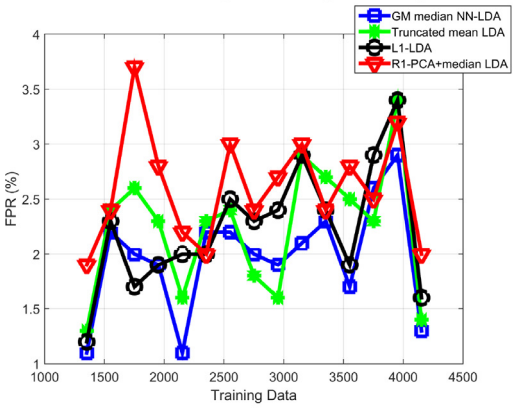
The first experiment consists in defining the adequate parameter p which increases the potential of GM Median NN-LDA. In theory, it is hard to achieve that. p is affected by several factors such as the total number of training samples, the number of total classes, the distribution of the samples. Consequently, the value of p often needs to be empirically determined. For instance, we consider p as 0.3, 0.5, 0.7 and 0.9. Fig. 6(a) and (b) show us that $p = 0.3$ is the value which obtains the highest average detection rate (DR) for KDDcup99 and NSL-KDD. Consequently, we set p to this value in the next experiments.

The next experiment compares GM Median NN-LDA to the following algorithms: LDA, median LDA, median NN-LDA. To avoid the (SSS) problem, PCA is used as the first stage of the

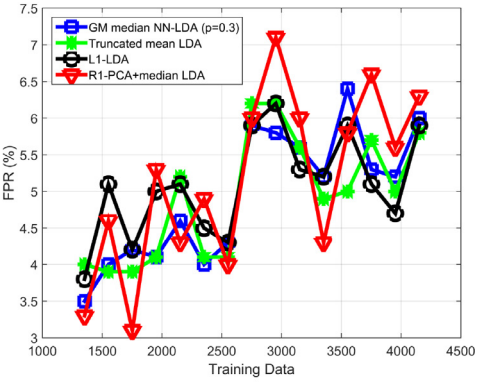


(a) : DR vs Training data using KDDcup99

(b) : DR vs Training data using NSL-KDD



(c) : FPR vs Training data using KDDcup99



(d) : FPR vs Training data using NSL-KDD

Fig. 8 – Comparison of GM median NN LDA with Truncated mean LDA, L1-LDA and R1-PCA+median LDA in term of DR and FPR.

Table 1 – Average individual attack detection rate(%) of GM Median NN-LDA, Median NN-LDA, LDA and median LDA.

The Database	The method	Normal	DOS	U2R	R2L	PROBE
KDDcup99	GM Median NN-LDA	97.9	**91.8**	**16.7**	**5.2**	**85.15**
	Median NN-LDA	98	90.7	14.1	5.2	84
	LDA	96.9	88.2	10.8	2.7	79.7
	median LDA	96.9	88	16.3	2.7	81.1
NSL-KDD	GM Median NN-LDA	93.8	**77.3**	**14**	**14**	**74.7**
	Median NN-LDA	94.1	77.2	12.5	14.2	73.3
	LDA	93.6	75	12.3	10.2	68
	median LDA	93.9	73.5	14.5	11.4	67.6

LDA, median LDA and median NN-LDA algorithms. Hence, these algorithms can also be noted as the PCA + LDA, PCA + median LDA, PCA + median NN-LDA. In PCA stage we selected 3 principal components. In LDA variant stage we have chosen 3 top features. Having said that, we increased the number of training data and we visualized it influence on DR and FPR of every method.

According to Fig. 7(a) and (b), it is observable that our approach takes the lead in attacks detection as the training data grows up. This is due to two reasons. The first one relies on median NN-LDA philosophy. More there are training samples, the easier the local structure around every class median can be captured and the boundaries of every class become more structured and separable. The second reason concerns the use of generalized mean ($p = 0.3$) which is less sensitive to outliers.

Fig. 7 (c) shows that in case of KDDcup99, GM Median NN-LDA produces the lowest false positive rate compared to the other methods. For NSL-KDD (Fig. 7(d)), the approach gives less than 6% of false positive rate. These facts prove the high ability of our approach to recognize the normal network instances regardless of training samples size.

In another simulation, we visualize the DOS, U2R, R2L and PROBE detection rate of the aforementioned LDA variants. We observe in Table 1 that the proposed approach takes the lead in identifying such attacks for the two datasets.

To further evaluate the performance of the approach, we compare it to other recent LDA methods such as Truncated

Table 2 – Average individual attack detection rate(%) of GM Median NN-LDA, Truncated Mean LDA, L1-LDA and R1-PCA+median LDA.

The Database	The method	Normal	DOS	U2R	R2L	PROBE
KDDcup99	GM Median NN-LDA	98.7	**92.8**	**18.4**	**4.2**	**84.9**
	Truncated Mean LDA	98.6	90.5	18.2	3.4	83.6
	L1-LDA	98.4	94.3	13.8	4	83.4
	R1-PCA+median LDA	98	95.4	15.8	1.2	83.8
NSL-KDD	GM Median NN-LDA	94.6	**74.6**	**15**	**12.4**	**74.2**
	Truncated Mean LDA	94.7	73.2	9	11.1	66.5
	L1-LDA	89.8	71.6	4.6	2.4	72.2
	R1-PCA+median LDA	93.9	71.8	16.4	0.4	75.1

mean LDA, L1-LDA and R1-PCA+median LDA. The three first principal components and the three first discriminant vectors are employed in the experiments. Fig. 8 exposes the obtained results. As it was done in the previous experiments, we varied the number of training samples from 1350 to 4150 and illustrate DR and FPR behaviors.

As regards the first dataset, it is clear from Fig. 8(a) that GM Median NN-LDA overcomes the three approaches once the size of training data is superior than 2100. In the other hand, Fig. 8(c) shows that the proposed approach give less than 3% of FPR. Concerning the individual attack detection. Table 2 shows that GM Median NN-LDA gives better U2R, R2L and PROBE identification rate. However, it is surpassed by L1-LDA and R1-PCA+median LDA in term of DOS detection rate.

In case we use NSL-KDD, it is shown from Fig. 8(b) that in term of the general DR, GM Median NN-LDA surpasses the recent LDA variants. Concerning FPR, Fig. 8(d) asserts that the approach still gives satisfactory results. From Table 2 we observe that the approach recognizes in better way DOS and R2L attacks in one hand with the values 74.6% and 12.4%. In the other hand, it is clear that GM Median NN-LDA gives competitive results in term of U2R and PROBE detection.

7. Conclusion

In context of network intrusion detection, this study proposed a robust Median NN-LDA using the generalized means. In order to deal with the outliers effect, the approach takes advantage of using the generalized means in within and between scatter matrices formulation. Many experiments were conducted on two popular Network data sets (KDDcup99 and NSL-KDD) to demonstrate that the proposed method performs better than many recent LDA approaches.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cose.2019.05.021.

REFERENCES

Abou-Moustafa KT, De La Torre F, Ferrie FP. Pareto models for discriminative multiclass linear dimensionality reduction. Pattern Recognit 2015;48(5):1863–77.

Bullen P. Handbook of means and their inequalities 2003;260.

Chen HT, Chang HW, Liu TL. Local discriminant embedding and its variants, 2. IEEE; 2005. p. 846–53.

Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.

Dai DQ, Yuen PC. Face recognition by regularized discriminant analysis. IEEE Trans Syst Man Cybern Part B (Cybern) 2007;37(4):1080–5.

Deng W, Hu J, Lu J, Guo J. Transform-invariant PCA: a unified approach to fully automatic facealignment, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 2014;36(6):1275–84.

Denning DE. An intrusion-detection model. IEEE Trans Softw Eng 1987(2):222–32.

Ding C, Zhou D, He X, Zha H. R 1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization. In: Proceedings of the 23rd international conference on machine learning. ACM; 2006. p. 281–8.

Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley & Sons; 2012.

Elkhadir Z, Chougdali K, Benattou M. Network intrusion detection system using PCA by lp-norm maximization based on conjugate gradient. Int Rev Comput Softw (IRECOS) 2016;11(1):64–71.

Elkhadir Z, Chougdali K, Benattou M. A median nearest neighbors LDA for anomaly network detection. In: Proceedings of the International Conference on Codes, Cryptology, and Information Security. Springer; 2017. p. 128–41.

Ghassabeh YA, Rudzicz F, Moghaddam HA. Fast incremental LDA feature extraction. Pattern Recognit 2015;48(6):1999–2012.

He R, Hu B, Yuan X, Zheng WS. Principal component analysis based on non-parametric maximum entropy. Neurocomputing 2010;73(10–12):1840–52.

He R, Hu BG, Zheng WS, Kong XW. Robust principal component analysis based on maximum correntropy criterion. IEEE Trans Image Process 2011;20(6):1485–94.

He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using laplacianfaces. IEEE Trans Pattern Anal Mach Intell 2005;27(3):328–40.

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

https://www.unb.ca/cic/datasets/nsl.html.

Ji Z, Pang Y, Yuan Y, Pan J. Relevance and irrelevance graph based marginal fisher analysis for image search reranking. Signal Process 2016;121:139–52.

Jolliffe I. Principal component analysis. Wiley Online Library; 2002.

Ke Q, Kanade T. Robust l/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming, 1. IEEE; 2005. p. 739–46.
Kwak N. Principal component analysis based on l1-norm maximization. IEEE Trans Pattern Anal Mach Intell 2008;30(9):1672–80.
Kwak N. Principal component analysis by l_{p}-norm maximization. IEEE Trans Cybern 2014;44(5):594–609.
Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol 2013;49(4):764–6.
Li H, Jiang T, Zhang K. Efficient and robust feature extraction by maximum margin criterion. In: Advances in neural information processing systems; 2004. p. 97–104.
Liang Z, Xia S, Zhou Y, Zhang L, Li Y. Feature extraction based on lp-norm generalized principal component analysis. Pattern Recognit Lett 2013;34(9):1037–45.
Lu J, Tan YP. Regularized locality preserving projections and its extensions for face recognition. IEEE Trans Syst Man Cybern Part B (Cybernetics) 2010;40(3):958–63.
Lu M, Huang JZ, Qian X. Sparse exponential family principal component analysis. Pattern Recognit 2016;60:681–91.
Oh J, Kwak N. Generalized mean for robust principal component analysis. Pattern Recognit 2016;54:116–27.
Oh J, Kwak N, Lee M, Choi CH. Generalized mean for feature extraction in one-class classification problems. Pattern Recognit 2013;46(12):3328–40.
Oh JH, Kwak N. Generalization of linear discriminant analysis using Lp-norm. Pattern Recognit Lett 2013;34(6):679–85.
Ren CX, Dao-Qing D, He X, Yan H. Sample weighting: an inherent approach for outlier suppressing discriminant analysis. IEEE Trans Knowl Data Eng 2015;27(11):3070–83.
Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J Mach Learn Res 2007;8(May):1027–61.
Sugiyama M, Idé T, Nakajima S, Sese J. Semi-supervised local fisher discriminant analysis for dimensionality reduction. Mach Learn 2010;78(1–2):35–61.
Wang H, Chen S, Hu Z, Zheng W. Locality-preserved maximum information projection. IEEE Trans Neural Netw 2008;19(4):571–85.
Wang S, Lu J, Gu X, Du H, Yang J. Semi-supervised linear discriminant analysis for dimension reduction and classification. Pattern Recognit 2016;57:179–89.
Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 2007;29(1):40–51.
Yang J, Zhang D, Yang JY. Median LDA: a robust feature extraction method for face recognition, 5. IEEE; 2006. p. 4208–13.
Zhang Y, Wang K, Gao M, Ouyang Z, Chen S. Lkm: a LDA-based k-means clustering algorithm for data analysis of intrusion detection in mobile sensor networks. Int J Distrib Sens Netw 2015;11(10):491910.
Zyad E, Khalid C, Mohammed B. Combination of R1-PCA and median LDA for anomaly network detection. In: Proceedings of the Intelligent Systems and Computer Vision (ISCV). IEEE; 2017a. p. 1–5.
Zyad E, Khalid C, Mohammed B. An effective network intrusion detection based on truncated mean LDA. In: Proceedings of the International Conference on Electrical and Information Technologies (ICEIT); 2017b. p. 1–5. doi:10.1109/EITech.2017.8255298.

Elkhadir Zyad is a Computer Science Doctor graduated from Faculty of science, Ibn Tofail University, Kenitra, Morocco. He obtained his Master degree in computer science in 2013 from the same Faculty. He is an IEEE member. His main research interest is to develop new feature extraction algorithms for pattern recognition problem such as network intrusion detection.

Mohamed Benattou is a Professor of Computer Science at the IBN TOFAIL University KNITRA where he directs the Computer Science and Telecommunication Laboratory. He has also held several positions in his French academic career: University of PAU, University of ORSAY Paris XI, 3IL and Xlim Laboratory. His research interests include distributed testing, secure testing, and software testing.