

# A Median Nearest Neighbors LDA for Anomaly Network Detection

Zyad Elkhadir<sup>1</sup>(✉), Khalid Chougali<sup>2</sup>, and Mohammed Benattou<sup>1</sup>

<sup>1</sup> LASTID Laboratory, Ibn Tofail University, Kenitra, Morocco  
zyad.elkhadir@gmail.com

<sup>2</sup> GEST Research Group, National School of Applied Sciences (ENSA),  
Ibn Tofail University, Kenitra, Morocco  
chougali@gmail.com

**Abstract.** The Linear Discriminant Analysis (LDA) is a powerful linear feature reduction technique. It often produces satisfactory results under two conditions. The first one requires that the global data structure and the local data structure must be coherent. The second concerns data classes distribution nature. It should be a Gaussian distribution. Nevertheless, in pattern recognition problems, especially network anomalies detection, these conditions are not always fulfilled. In this paper, we propose an improved LDA algorithm, the median nearest neighbors LDA (median NN-LDA), which performs well without satisfying the above two conditions. Our approach can effectively get the local structure of data by working with samples that are near to the median of every data class. The further samples will be essential for preserving the global structure of every class. Extensive experiments on two well known datasets namely KDDcup99 and NSL-KDD show that the proposed approach can achieve a promising attack identification accuracy.

**Keywords:** LDA · median NN-LDA · Network anomaly detection · NSL-KDD · KDDcup99

## 1 Introduction

The linear discriminant analysis (LDA) [1] is a family of techniques whose role is dimensionality reduction and feature extraction. Fishers LDA is one of the most known LDA methods. It has been used successfully in a variety of pattern recognition problems including network anomalies detection [2–4]. The key procedure behind Fishers LDA or LDA is to employ 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.

In literature, many works have been proposed to ameliorate the performance and the accuracy of the classical LDA. These works can be generally divided into

two categories. The first category tries to solve the small sample size (SSS) problem, which always happens when the data dimension is greater than the number of training samples. As noticed in previous contributions, to overcome the SSS problem, direct linear discriminant analysis (Direct LDA) [5] eliminates the null space of the inter-class scatter matrix as a first step. After that, it extracts the discriminant information from the null space of the intra-class scatter matrix. In the same way, Null space LDA [6] exploited the valuable discriminant vectors of the null space of the intra-class scatter matrix with the help of PCA [7]. These vectors are used rather than the eigenvectors of the classical LDA. The authors of the last method also demonstrated that the extracted vectors are equivalent to the optimal LDA discriminant vectors obtained in the original space.

In [8] we can see an exponential discriminant analysis algorithm that derive the most discriminant information which exists in the intra-class scatter matrix's null space. However, the procedures employed by the aforementioned algorithms destroy a big part of discriminant information essential for classification. Another technique to overcome the (SSS) problem is presented in [9]. It employed an optimization criterion which used a generalized singular value decomposition. This technique is operational regardless of whether the dimension of data is greater than the number of training samples. Alternatively, an ensemble learning framework was developed by Wang and Tang [10] in order to preserve the significant discriminant information by random sampling on feature vectors and training samples. In [11], three LDA approaches were proposed to solve the SSS problem: regularized discriminant analysis [12], discriminant common vectors [13], and Maximum Margin Criterion (MMC) [14]. Another famous approach to address the SSS consists in using PCA + LDA to get the discriminant features (i.e., apply PCA on data before LDA). Nevertheless, this method may lose valuable discriminant information in PCAs stage.

The second part of works deals with the incremental versions of the LDA. This kind of LDA is very useful for online learning tasks. One of their main advantages is that the feature extraction method does not need to save the entire data matrix in the memory. In [15], QR decomposition with a LDA-based incremental algorithm were proposed. In [16], the authors developed many incremental LDAs which have a common point. The algorithms have to update in every step the between-class and within-class scatter matrices. Another incremental LDA is presented in [17]. Here, the authors showcase a good mechanism to update the scatter matrices. Besides the above two kind of improvements of LDA, there are also some LDA-based algorithms such as R1 LDA [18], L1 LDA [19], Median LDA [20] and pseudo LDA [21].

Unfortunately, all these aforementioned LDA methods pay more attention to the global structure of classes. As a result, the produced discriminant vectors are often skewed. Before going through the explanation of this fact, we give an overview of class distribution types. In general, there is two kind of complementary distributions. One is local and the other is global. The first one represents a portion of samples that defines in a certain manner the real distribution nature of every class. In the other hand, the global distribution determines the class

boundaries and helps us to separate as much as possible the classes. However, in reality, the last distribution is in most of cases not Gaussian and has a more complex structure. In addition, it is often incoherent with the first type of distribution. All these assumptions lead to an inaccurate discriminant vectors.

In order to address this matter, previous works [22–24] exploited local information to obtain optimal discriminant vectors. Nonetheless, in these works, it is necessary to calculate a matrix where each of its element is a distance between two data samples, in addition, we have to do an eigen decomposition of a huge matrix generated by the entire training set. For network intrusion detection field it will be a time consuming and even an infeasible task. As a result, it is difficult to implement these approaches.

In this paper, to deal with the drawback of the global LDA, we propose a kind of local LDA namely Median Nearest Neighbors LDA. The method takes into account also preserving the global structure. Our approach consists of two parts. The first part is to find a proper number of nearest neighbors to the median of every class training set. The determined nearest neighbors will be used to compute the within-class scatter matrix. In the second part, the rest of samples which are further from the median will determine the between-class scatter matrix.

The rest of this paper is organized as follows. In Sect. 2, we outline the classical LDA. Section 3 presents in details the proposed approach. Section 4 introduces the two well known network datasets KDDcup99 and NSL-KDD. In Sect. 5 we give the experimental results and illustrate the effectiveness of the algorithm and compare it to some of the above LDA approaches. Finally, Sect. 6 offers our conclusions.

## 2 Linear Discriminant Analysis

The conventional LDA aims to reduce dimensionality while keeping the maximum of class-discriminatory information. This operation is realized by projecting original data onto a lower dimensional space with taking into account maximizing separation of different classes on the one hand, and minimizing dispersion of samples of the same class on the other hand. Mathematically speaking, suppose we have a data matrix  $X = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}$  composed of  $n$  samples, our purpose is to 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 rewritten as  $X = [X_1, \dots, X_k]$  such that  $k$  is the number of classes and  $X_i \in \mathbb{R}^{d \times n_i}$  represents samples of the  $i$ th class,  $n_i$  is the sample size of the  $i$ th class and  $\sum_{i=1}^k n_i = n$ . LDA operates on three important matrices namely within-class, between-class and total-scatter matrices which are defined as follows:

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

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

$$S_t = (1/n) \sum_{i=1}^n (x_i - c)(x_i - c)^T \quad (3)$$

$c_i$  is the mean of the  $i$ th class, and  $c$  is the general mean. It can be proved that  $S_t = S_w + S_b$  [1]. It follows from (1) and (2) that:

$$\text{trace}(S_w) = (1/n) \sum_{i=1}^k \sum_{x \in X_i} \|x - c_i\|^2 \quad (4)$$

$$\text{trace}(S_b) = (1/n) \sum_{i=1}^k n_i \|c_i - c\|^2 \quad (5)$$

The trace of  $S_w$  gives us an idea on how every sample is close to its 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 three scatter matrices become:

$$\begin{aligned} S_w^l &= G^T S_w G \\ S_b^l &= G^T S_b G \\ S_t^l &= G^T S_t G \end{aligned}$$

The optimal projection matrix can be gained by maximizing the following objective function:

$$G = \arg \max \frac{\text{trace}(S_b)}{\text{trace}(S_w)} \quad (6)$$

When  $S_w$  is invertible, the solutions to (6) can be obtained by performing the following generalized eigenvalue decomposition:

$$S_w^{-1} S_b g_i = \lambda_i g_i \quad (7)$$

where  $G = [g_1, \dots, g_l]$ .

Setting aside the famous (SSS) problem, LDA suffers from another matter. It uses the global structure information of the total training samples to determine the linear discriminant vectors. In general, the use of these vectors to extract features from the samples may lead to erroneous classification. The potential reason behind this phenomenon seems to be that the global distribution of the data does not represent the real distribution nature of every class. In other words, the global distribution is not always consistent with the local distribution. Moreover, the non Gaussian nature of data might cause a nonlinear boundaries between the classes. So it becomes difficult to use global linear discriminant vectors to separate the data.

### 3 The Proposed Method

To overcome the aforementioned LDA drawbacks, we propose to exploit the local distribution of every class. To do that we were based on the concept of median. In probability theory and statistics, the median is defined as a sample that separates the higher half of a probability distribution from the lower half. It is the middle value in a distribution, above and below which lie an equal number of samples. From this assumption, we observe that 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 we can assimilate the further samples to the global distribution, since they exist naturally in the boundaries of the class and facilitate the separation of classes. With this concept we dissociate the two distributions. Therefore, we resolve the matter of distribution's consistency.

Our approach (median NN-LDA) also performs well even if the data is not Gaussian or has nonlinear boundaries. Since it can extract the global structures of the data through determining the samples which are far from the median, the method can obtain a number of local linear discriminant vectors which approximate the nonlinear boundary between the classes.

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

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

Let  $X_i^b = [x_{p+1}, \dots, x_{n_i}] \in \mathbb{R}^{d \times (n_i - p)}$  contains the  $n_i - p$  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:

$$S'_w = (1/p) \sum_{i=1}^k \sum_{x \in X_i^w} (x - c_i^w)(x - c_i^w)^T \quad (8)$$

$$S'_b = (1/p) \sum_{i=1}^k (c_i^w - c)(c_i^w - c)^T \quad (9)$$

Where  $c_i^w$  is the mean of  $X_i^w$ ,  $c_i^b$  is the mean of  $X_i^b$  and  $c = \frac{1}{k} \sum_{i=1}^k (c_i^b)$  is the general mean.

As a consequence, Eqs. (4) and (5) will be replaced by:

$$trace(S'_w) = (1/p) \sum_{i=1}^k \sum_{x \in X_i^w} \|x - c_i^w\|^2 \quad (10)$$

$$trace(S'_b) = (1/p) \sum_{i=1}^k n_i \|c_i^w - c\|^2 \quad (11)$$

We obtain the discriminant vectors by maximizing the following objective function:

$$G' = \arg \max \frac{\text{trace}(S'_b)}{\text{trace}(S'_w)} \quad (12)$$

The solution can be reached by performing:

$$(S'_w)^{-1} S'_b g'_i = \lambda'_i g'_i \quad (13)$$

where  $G' = [g'_1, \dots, g'_l]$ .

In order to deal with the singularity problem, we propose to apply an intermediate dimensionality reduction stage, such as principal component analysis (PCA) [7] to reduce the data dimensionality before applying median NN-LDA.

## 4 The Simulated Databases and Its Transformation

### 4.1 KDDcup99

The KDDcup99 [25] 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 [26] is a new version of KDDcup99 dataset. This dataset has some advantages over the old one and has addressed some of its 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.

## 4.3 Transformation Process

In order to successfully apply the approach on the datasets, as a crucial step, we have converted all the discrete attributes values of the datasets to continuous values. To accomplish that, we applied the following procedure: every discrete attribute  $i$  which takes  $k$  different values will be represented as  $k$  coordinates composed of ones and zeros. For example, we know that the protocol type attribute has three values tcp, udp or icmp. According to the procedure, all these values will be transformed to the corresponding coordinates (1, 0, 0), (0, 1, 0) or (0, 0, 1).

## 5 Experiments and Discussion

In this section, in order to demonstrate the effectiveness of the proposed method, we conduct a series of experiments with KDDcup99 and NSL-KDD. Meanwhile, we also compare median NN-LDA performance with LDA, direct LDA, null space LDA, R1 LDA, pseudo LDA in an all-round way.

We can employ the following measures to evaluate these methods:

$$DR = \frac{TP}{TP + FN} \times 100 \quad (14)$$

$$FPR = \frac{FP}{FP + TN} \times 100 \quad (15)$$

In network security jargon, (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 normal instances classified as normal. Therefore, the most performant feature extraction method, is the one which produces a high DR and a low FPR.

In our experiments, we varied the size of training samples and kept test dataset intact with the following composition (100 normal data, 100 DOS data, 50 U2R data, 100 R2L data, and 100 PROBE). To reduce the variation of the detection rate (DR), we adopt the mean of twenty runs. Since our aim is to evaluate the efficacy of feature extraction method, we use a simple classifier, the nearest neighbor classifier.

The first experiment consists in defining the adequate number of samples  $p$  which represent the local structure of every class. In theory, it is difficult to do that. The most suitable  $p$  is affected by several factors such as the total number

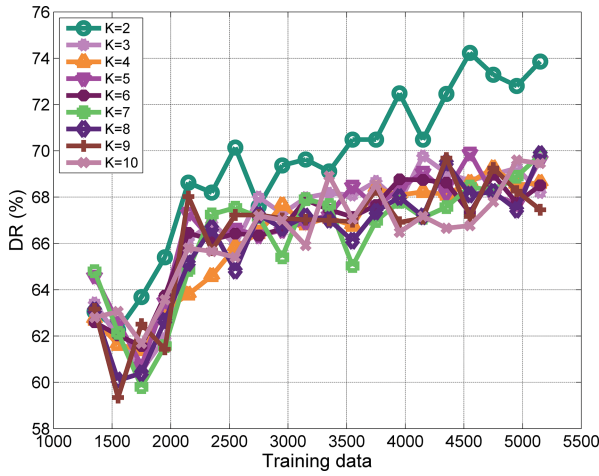


Fig. 1. Detection rate of different K for KDDcup99

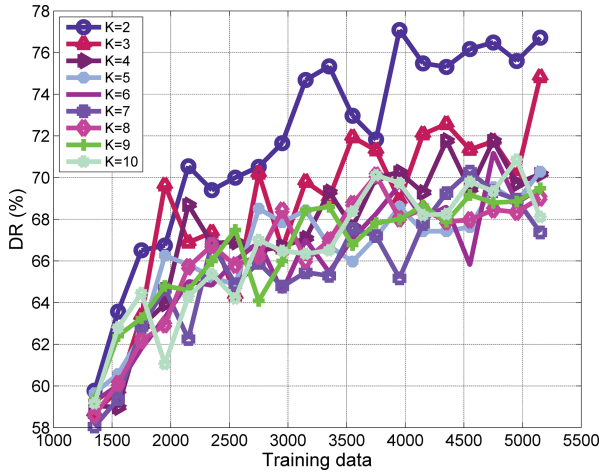


Fig. 2. Detection rate of different K for NSL-KDD



of training samples, the number of total classes, the distribution of the samples. Therefore, the value of  $p$  often needs to be empirically determined. For instance, we consider  $p$  as  $\frac{n_i}{K}$  and we varied  $K$  from 2 to 10. Figures 1 and 2 show us that  $p = \frac{n_i}{2}$  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.

In the second experiment we compare our proposed method to the following algorithms: LDA, median LDA, null space LDA, Direct LDA and pseudo LDA. To avoid the (SSS) problem, PCA is used as the first stage of the LDA,

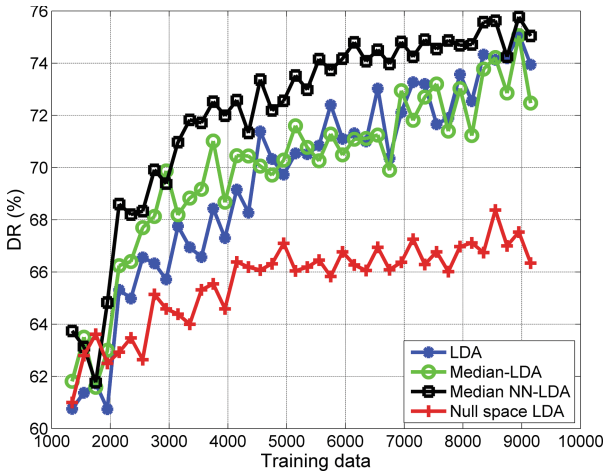


Fig. 3. Training data vs. detection rate for KDDcup99

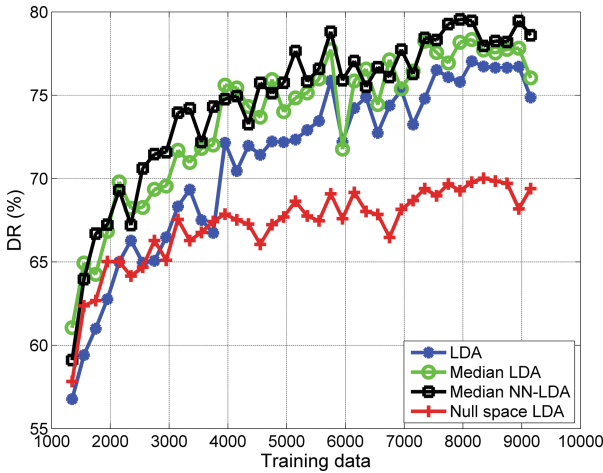


Fig. 4. Training data vs. detection rate for NSL-KDD

median LDA and median NN-LDA algorithms. Hence, these algorithms can also be viewed as the PCA + LDA, PCA + median LDA, PCA + median NN-LDA. We have chosen 3 principal components in the first stage of these methods. In the second stage we have chosen 3top features. The rest of LDA algorithms exploit the 4 top discriminant vectors. Having said that, we increased the number of training data and we visualized it influence on DR and FPR of every method.

Figures 3, 4, 5 and 6 illustrate the results we found when we compare our approach to LDA, median LDA and null space LDA for the two datasets. According to the first two figures, we observe that our approach takes the lead in attacks

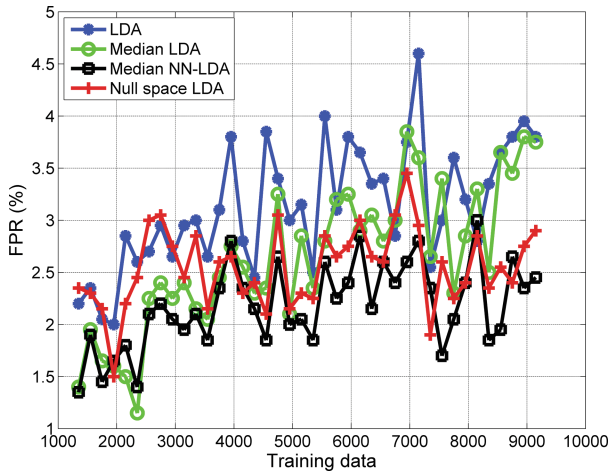


Fig. 5. Training data vs. FPR for KDDcup99

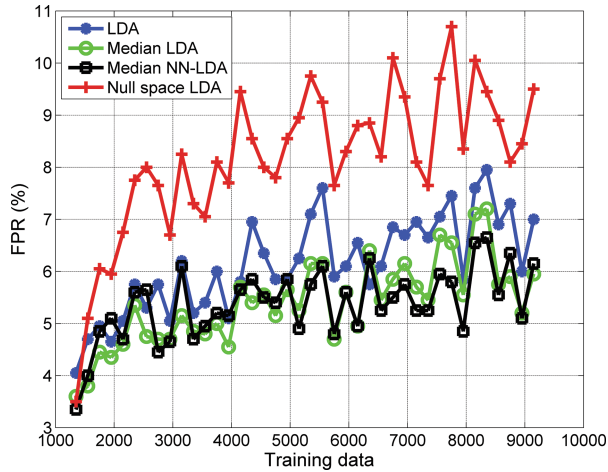


Fig. 6. Training data vs. FPR for NSL-KDD

detection as the training data grows up. The reason behind this phenomenon seems to be that more we have training samples, the easier the local structure around every class median can be captured. In addition, when we increase the number of training samples, the boundaries of every class become more structured and separable. This truth helps as much as possible in preserving the global distribution. The rest of figures depict the relationship between training samples and FPR. It is clear that median NN-LDA produces the lowest false positive rate compared to the other methods. This fact proves the high ability of our approach to recognize the normal network instances regardless of training samples size.

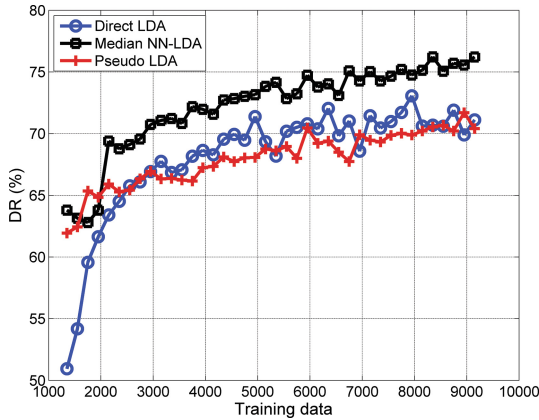


Fig. 7. Training data vs. detection rate for KDDcup99

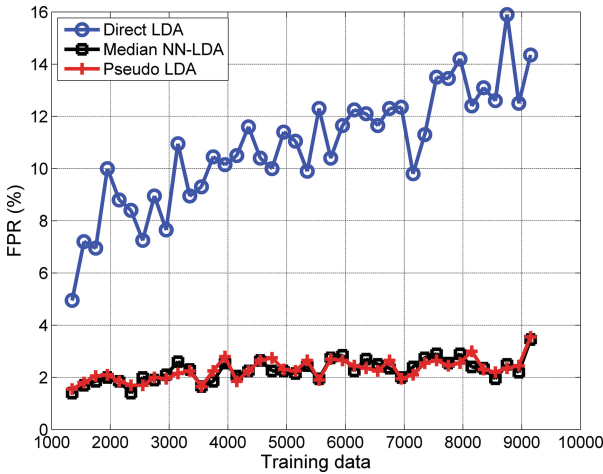


Fig. 8. Training data vs. FPR for KDDcup99

To further evaluate the performance of our approach, we compare it to other LDA methods such as Direct LDA and pseudo LDA. Figures 7, 8, 9 and 10 expose the obtained results while using KDDcup99 and NSL-KDD. As we have done in the previous experiments, we varied the number of training samples from 1350 to 9150 and illustrate DR and FPR behaviors.

As regards the first dataset, we observe from Fig. 7 that median NN-LDA overcomes the two approaches once the size of training data is superior than 2000. In the other hand, Fig. 8 shows that Pseudo LDA and the proposed approach give the fewest number of false positives.

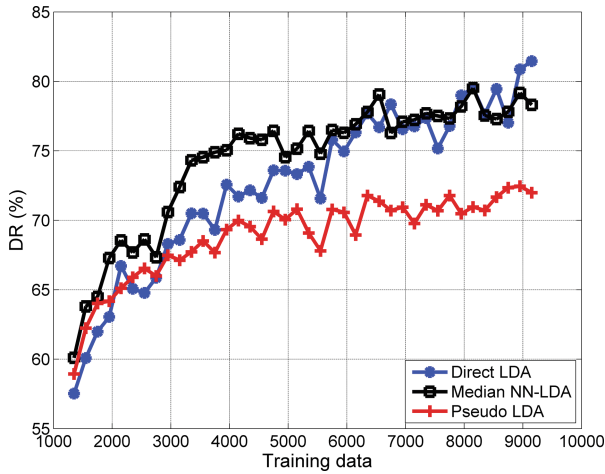


Fig. 9. Training data vs. detection rate for NSL-KDD

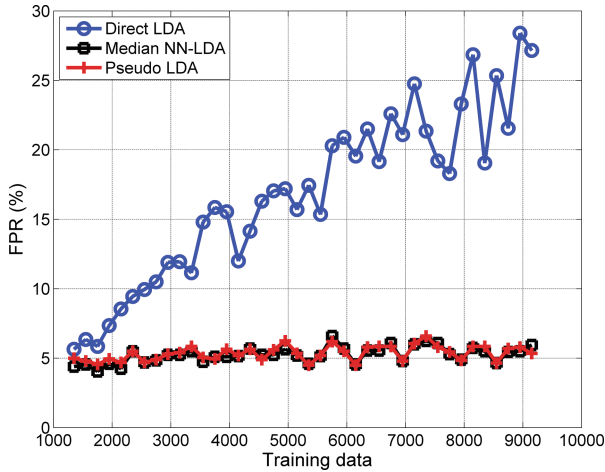


Fig. 10. Training data vs. FPR for NSL-KDD

In case we use NSL-KDD, it is shown from Fig. 9 that in term of DR, median NN-LDA surpasses Direct LDA and Pseudo LDA when the training dataset size is less than 8000. Once this value is exceeded, Direct LDA starts to compete with median NN-LDA. Concerning FPR, Fig. 10 asserts that our approach still gives satisfactory results.

## 6 Conclusion

In this paper, a novel feature extraction method called median NN-LDA is proposed. In this LDA approach we exploit the median of every class to compute the within and between scatter matrices. There are two advantages of median NN-LDA, one is that it preserves the local and the global distributions, the other is its insensitivity to non Gaussian data. Therefore, the proposed method is more robust than traditional linear discriminant analysis. We conduct the experiments on two popular Network data sets (KDDcup99 and NSL-KDD), using many LDA approaches. The experimental results indicate that the proposed method has a promising performance.

## References

1. Fukunaga, R.: *Statistical Pattern Recognition*. Academic Press, New York (1990)
2. Thapngam, T., Yu, S., Zhou, W.: DDoS discrimination by linear discriminant analysis (LDA). In: 2012 International Conference on Computing, Networking and Communications (ICNC), pp. 532–536. IEEE (2012)
3. An, W., Liang, M.: A new intrusion detection method based on SVM with minimum within-class scatter. *Secur. Commun. Netw.* **6**(9), 1064–1074 (2013)
4. Subba, B., Biswas, S., Karmakar, S.: Intrusion detection systems using linear discriminant analysis and logistic regression. In: 2015 Annual IEEE India Conference (INDICON), pp. 1–6. IEEE (2015)
5. Yu, H., Yang, J.: A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recogn.* **34**(10), 2067–2070 (2001)
6. Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C., Yu, G.J.: A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recogn.* **33**(10), 1713–1726 (2000)
7. Jolliffe, I.: *Principal Component Analysis*. Wiley Online Library (2002)
8. Zhang, T., Fang, B., Tang, Y.Y., Shang, Z., Xu, B.: Generalized discriminant analysis: a matrix exponential approach. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **40**(1), 186–197 (2010)
9. Ye, J., Janardan, R., Park, C.H., Park, H.: An optimization criterion for generalized discriminant analysis on under sampled problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(8), 982–994 (2004)
10. Wang, X., Tang, X.: Random sampling for subspace face recognition. *Int. J. Comput. Vis.* **70**(1), 91–104 (2006)
11. Liu, J., Chen, S., Tan, X.: A study on three linear discriminant analysis based methods in small sample size problem. *Pattern Recogn.* **41**(1), 102–116 (2008)
12. Dai, D.Q., Yuen, P.C.: Regularized discriminant analysis and its application to face recognition. *Pattern Recogn.* **36**(3), 845–847 (2003)

13. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(1), 4–13 (2005)
14. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans. Neural Netw.* **17**(1), 157–165 (2006)
15. Ye, J., Li, Q., Xiong, H., Park, H., Janardan, R., Kumar, V.: IDR/QR: an incremental dimension reduction algorithm via QR decomposition. *IEEE Trans. Knowl. Data Eng.* **17**(9), 1208–1222 (2005)
16. Pang, S., Ozawa, S., Kasabov, N.: Incremental linear discriminant analysis for classification of data streams. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **35**(5), 905–914 (2005)
17. Kim, T.K., Wong, S.F., Stenger, B., Kittler, J., Cipolla, R.: Incremental linear discriminant analysis using sufficient spanning set approximations. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007*, pp. 1–8. IEEE (2007)
18. Li, X., Hu, W., Wang, H., Zhang, Z.: Linear discriminant analysis using rotational invariant L1 norm. *Neurocomputing* **73**(13), 2571–2579 (2010)
19. Wang, H., Lu, X., Hu, Z., Zheng, W.: Fisher discriminant analysis with L1-norm. *IEEE Trans. Cybern.* **44**(6), 828–842 (2014)
20. Yang, J., Zhang, D., Yang, J.Y.: Median LDA: a robust feature extraction method for face recognition. In: *IEEE International Conference on Systems, Man and Cybernetics, SMC 2006*, vol. 5, pp. 4208–4213. IEEE (2006)
21. Golub, G.H., Van Loan, C.F.: *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Hopkins University Press, Baltimore (1996)
22. Sugiyama, M., Idé, T., Nakajima, S., Sese, J.: Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Mach. Learn.* **78**(1–2), 35–61 (2010)
23. Chen, H.T., Chang, H.W., Liu, T.L.: Local discriminant embedding and its variants. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 2, pp. 846–853. IEEE (2005)
24. Wang, H., Chen, S., Hu, Z., Zheng, W.: Locality-preserved maximum information projection. *IEEE Trans. Neural Netw.* **19**(4), 571–585 (2008)
25. <http://kdd.ics.uci.edu/databases/kddcup99/>
26. <http://nsl.cs.unb.ca/NSL-KDD/>