Improving network intrusion detection using fuzzy LDA

ELKHADIR Zyad[1], BRITEL Merieme[1], CHOUGDALI Khalid[3], and BENATTOU Mohammed[1]

[1]LASTID Research Laboratory, Ibn Tofail University
[3]GEST Research group, National School of Applied Sciences (ENSA), Ibn Tofail University

***Abstract*—This paper proposes to use a feature extraction method called fuzzy LDA to ameliorate the detection rate of a network intrusion detection system (IDS). In this method, a membership degree matrix is calculated using Fuzzy c-means (FCM), then the membership degree is incorporated into the definition of the between-class scatter matrix and within class scatter matrix to get the fuzzy between-class scatter matrix and fuzzy within-class scatter matrix. Experimental results on KDDcup99 and NSL-KDD show that the proposed approach outperforms the classical LDA.**

***Keywords*—LDA, fuzzy theory, Network Anomaly Detection, NSL-KDD, KDDcup99.**

I. INTRODUCTION

Nowadays, many communication network tools have been developed, that what leads to the birth of sophisticated attacks. To alleviate that, a Numerous techniques like cryptography, honey pots and firewalls have been suggested. Nevertheless, these techniques show a lot of weaknesses, as a consequence, attackers in most of cases find their way around them to gain unauthorized access to the network and launch attacks. In addition, they fail in detecting inside attacks, where the attacker is a legitimate member of the network. To overcome these shortcomings, intrusion detection system(IDS) has been suggested in the literature [1]. It is seen as a complementary protection tool with other preventive mechanisms like authentication, data encryption.

IDS can be a signature-based IDS or an anomaly based IDS. The first category relies on a database of known attack signatures to identify malicious threats. It produces an alarm whenever there is a correspondence between the malicious network activities and one or more stored signatures. This kind of IDS has high detection rate against known attacks. Nevertheless, it is not able to detect new attacks. To overcome this limitation, frequent and expensive updates to the signature database are required. On the other hand, anomaly based IDS constructs a model with the help of network features. in this view, the attack represents any deviation of traffic patterns from the model. The main advantage of this category is it ability to identify zero day attacks. However, anomaly based IDS is emphasized by high false positive (FP) rate and low detection rate (DR).

To solve this issue, many papers such [2] [3] [4] exploited a feature extraction method called linear discriminant analysis (LDA) [5] and show a remarkable results. In this method, we project the network connections into a linearly independent discriminant vectors in order to separate as much as possible the normal and malicious classes. Consequently, this step provides the model with an important discrimination power which leads to a DR amelioration and FPR minimization.

However, LDA does not take into account overlapping classes in it formulation, a fact that decreases it efficiency. In such cases, one may try non-linear discriminant analysis methods, Nevertheless, the latter are in general complicated, and high time consuming. As an alternative to that, some measures should be adopted to reduce the influence of those overlapping data points. Fuzzy set theory may be useful to solve the above problem.

Fuzzy set theory was developed in 1965 by Zadeh. It defines a degree of class membership between zero and one instead taking values 0 or 1 as in ordinary set theory. From the point of view of applications, such an idea is helpful, since network data sets are usually ill-defined classes. For this reason, a fuzzy linear discriminant analysis is introduced in the present paper. The results so far indicate that Fuzzy LDA is an improvement over LDA, and that may provide more information about the structure of data sets being studied.

The rest of this paper is organized as follows. In Section II, we outline the theory of LDA. Then fuzzy LDA is introduced in Section III. Section IV describes the two well known network datasets KDDcup99 and NSL-KDD. Section V provides the experimental results and illustrates the effectiveness of the algorithm by comparing it to LDA approach. Finally, Section VI offers our conclusions.

II. LINEAR DISCRIMINANT ANALYSIS

The goal of LDA is finding a projection matrix G such that the Fisher criterion is maximized after the projection of samples. Suppose X is composed of k classes, $[X_1, .., X_k]$. Every X_i contains n_i samples. The between-class and within-class scatter matrices S_b and S_w, are defined by

$$S_w = (1/n) \sum_{i=1}^{k} \sum_{i \in X_i} (x - x'_i)(x - x'_i)^T \quad (1)$$

$$S_b = (1/n) \sum_{i=1}^{k} (x'_i - x')(x'_i - x')^T \quad (2)$$

x'_i is the mean of the ith class, and x' is the general mean. They are defined as follow:

$$x'_i = \frac{1}{n_i} \sum_{i \in X_i} (x_i) \tag{3}$$

and

$$x' = \frac{1}{n} \sum_{i=1}^{k} \sum_{i \in X_i} (x_i) \tag{4}$$

The Fisher criterion is defined by

$$G = \arg\max \frac{G^T S_b G}{G^T S_w G} \tag{5}$$

When S_w is not singular, the solutions to (5) can be obtained by solving the following equation:

$$S_w^{-1} S_b g_i = \lambda_i g_i \tag{6}$$

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

From the description above, it can be observed that the data samples belonging to the same class have the same weights in calculating the mean vector and the scatter matrices. Such a mechanism leads to a very weak ability to confront the influence of a few overlapping samples. The following Fuzzy LDA is designed to alleviate this problem.

III. FUZZY LDA FORMULATION

In the classical LDA, all data samples are treated in the same level. While in Fuzzy LDA, different samples have various degree of importance. Overlapping samples contribute much less than other samples. Moreover, their effect changes with the degree of overlapping. Thus, the data set X cannot be directly used in LDA, so the first thing is to fuzzify the data set and get memberships and centroids. To do that we run FCM algorithm on the data til termination to obtain memberships matrix U and centroids V. The steps are described in Algorithm 1:

After that, we will exploit U and V in the calculation of the fuzzy scatter matrices S_{fw} and S_{fb} with the following formulas:

$$S_{fw} = (1/n) \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^m (x_j - v_i)(x_j - v_i)^T \tag{7}$$

$$S_{fb} = (1/n) \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^m (v_i - x')(v_i - x')^T \tag{8}$$

The new Fisher criterion will be reformulated by

$$G' = \arg\max \frac{G'^T S_{fb} G'}{G'^T S_{fw} G'} \tag{9}$$

The solutions to the above problem is reached by:

$$(S_{fw})^{-1} (S_{fb}) g'_i = \lambda_i g'_i \tag{10}$$

Algorithm 1: Fuzzy C-means

Input : X the training data, k the number of classes and m the degree of fuzzification

Output: U and V

initialization: $b \longleftarrow 0$ (b indicates the current iteration);
repeat
  Compute centroids: $v_i^{b+1} = \frac{\sum_{i=1}^{n} (u_{ij}^b)^m x_j}{u_{ij}^b}$;
  Compute distances: $d_{ij}^2 = ||x_j - v_i^{b+1}||$;
  S.t $j = 1, .., n$ and $i = 1, .., k$;
  Update memberships matrix U;
  if $d_{ij}^2 > 0$ **then**
    $u_{ij}^{b+1} = \frac{1}{\sum_{i=1}^{k} (d_{ij}/d_{kj})^{2/m-1}}$;
  else
    if $u_{ij}^{b+1} \in [0,1] and \sum_{i=1}^{k} u_{ij}^{b+1} = 1$ **then**
      $u_{ij}^{b+1} = 0$;
    end
  end
  $b \longleftarrow b + 1$;
until $||U^{b+1} - U^b|| \leq \delta$;

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

In order to deal with the singularity problem, we propose to apply an intermediate dimensionality reduction stage, such as principal component analysis (PCA) [6] to reduce the data dimensionality before applying Fuzzy LDA.

IV. THE SIMULATED DATABASES

A. KDDcup99

The objective of 1999 KDD intrusion detection contest is to create a standard dataset [7] to evaluate research in intrusion detection. The dataset is prepared and managed by DARPA Intrusion Detection Evaluation Program. It is composed of many TCPdump raws, captured during nine weeks.

The first seven weeks were devoted to create training data. The latter represents four gigabytes of compressed binary TCP dump data, equivalent to five million connection records. Similarly, in last two weeks, the program captured around two million connection records and considered it as testing data. The KDD dataset was employed in the UCI KDD1999 competition whose goal is developing intrusion detection system models. the attacks simulated in this competition fall into four main categories: DOS, R2L, U2R, PROBE. In the first category an attacker tries to prevent legitimate users accessing or consume a service via back, land, Neptune, pod Smurf and teardrop. In R2L, the attacker tries to gain access to the victim system by compromising the security via password guessing or breaking. To perform U2R, the intruder tries to access super users (administrators) privileges via Buffer overflow attack. The last type of attack consists in gaining information about the victim machine by checking vulnerability on the victim machine. e.g., Port scanning.

The KDD Cup99 dataset is available in three different files such as KDD Full Dataset which contains 4898431 instances,

KDD Cup 10% dataset which contains 494021 instances, KDD Corrected dataset which contains 311029 instances. In this paper, training data are taken from KDD Cup 10% and testing data from KDD Corrected dataset.

Each sample of the dataset is a connection between two network hosts according to network protocols. It is described by 41 attributes. 38 of them are continuous or discrete numerical attributes, the other are categorical attributes. Each sample is labeled as either normal or one specific attack. The dataset contains 23 class labels out of which 1 is normal and remaining 22 are different attacks. The total 22 attacks fall into four categories as forth-mentioned attacks.

B. NSL-KDD

NSL-KDD is a data set [8] proposed to solve some of the shortcomings of the KDD'99 data set discussed in [9] . To summarize, 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

The advantages of Fuzzy LDA over LDA for network intrusion data sets are verified by the two data sets mentioned above.

To estimate the accuracy of these methods we employ two factors:

$$DR = \frac{TP}{TP + FN} \times 100 \tag{11}$$

$$FPR = \frac{FP}{FP + TN} \times 100 \tag{12}$$

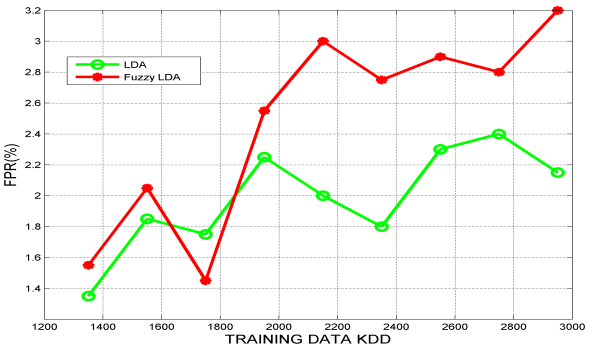
(DR) and (FPR) are Detection Rate and False Positive Rate. True positives (TP) refer to attacks correctly predicted. False negatives (FN) represent intrusions classified as normal instances, false positive (FP) are normal instances wrongly classified, and true negatives (TN) are normal instances classified as normal. The best feature extraction method will be the one which improves DR as much as possible and tries to minimize FPR.

In the experiments, we increase the number of training samples and fixed test dataset with the composition (100 normal data, 100 DOS data, 50 U2R data, 100 R2L data, and 100 PROBE). To have a realistic DR and FPR, the operation of sample selection was done randomly for twenty times. Then DR and FPR took the average. For the sake of simplicity we use the nearest neighbor classifier.

Figs. 1 and 2 show the results we found when we compare our approach to LDA on KDDcup99. According to the first figure, we observe that fuzzy LDA overcomes LDA permanently in attack detection. That is something trivial, because more there are training samples more the effect of

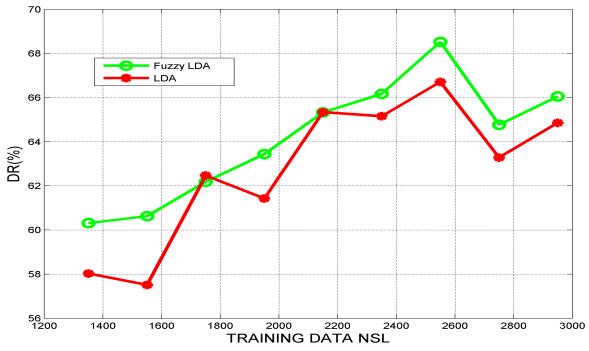
Fig. 1: DR of fuzzy LDA and LDA on KDDcup99

Fig. 2: FPR of fuzzy LDA and LDA on KDDcup99



overlapping samples is visible. Since fuzzy LDA gives degrees of membership to these samples it will be more robust than LDA. Fig. 2 illustrates the comparison in term of FPR. We can observe that the proposed method is competitive with LDA when training samples are less than 2000. Once this value is exceeded, fuzzy LDA shows high FPR.

Fig. 3: DR of fuzzy LDA and LDA on NSL-KDD



On NSL-KDD, we observe from Figs. 3 and 4 that fuzzy LDA show some improvement over classical LDA in DR. However, it still produce important FPR. One reason behind this weakness may comes from the effect of outliers. Effectively, in the fuzzification part where FCM clustering algorithm

Fig. 4: FPR of fuzzy LDA and LDA on NSL-KDD

is employed, we compute distances d using the euclidean norm, a norm which has been known to be sensitive to outliers.

VI. Conclusion

LDA does not take into account overlapping classes in it formulation, a fact that decreases it efficiency in many applications, in particular intrusion detection. To alleviate that, in this paper we propose to combine Fuzzy set theory with LDA. we perform FCM algorithm to fuzzify the data set and get memberships and centroids. After that we worked with the fuzzy between-class scatter matrix and fuzzy within-class scatter matrix instead of the classical ones. Experiments on KDDcup99 and NSL-KDD demonstrate the effectiveness of the proposed model.

References

[1] A. Abduvaliyev, A.-S. K. Pathan, J. Zhou, R. Roman, and W.-C. Wong, "On the vital areas of intrusion detection systems in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1223–1237, 2013.

[2] T. Thapngam, S. Yu, and W. Zhou, "Ddos discrimination by linear discriminant analysis (lda)," in *Computing, Networking and Communications (ICNC), 2012 International Conference on*. IEEE, 2012, pp. 532–536.

[3] B. Subba, S. Biswas, and S. Karmakar, "Intrusion detection systems using linear discriminant analysis and logistic regression," in *2015 Annual IEEE India Conference (INDICON)*. IEEE, 2015, pp. 1–6.

[4] Z. Tan, A. Jamdagni, X. He, and P. Nanda, "Network intrusion detection based on lda for payload feature selection," in *2010 IEEE Globecom Workshops*. IEEE, 2010, pp. 1545–1549.

[5] R. FUKUNAGA, "Statistical pattern recognition," 1990.

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

[7] [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/

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

[9] 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.