The CSI Journal on
Computer Science and Engineering
Vol. 15, No. 1, 2017
Pages 22-30
Regular Paper

Email Spam Detection Using Linear Discriminant Analysis Based on Clustering

Maryam Imani[1] Gholam Ali Montazer[2]

[1]Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
[2]Faculty of Information Technology Engineering, Tarbiat Modares University, Tehran, Iran

Abstract

The high volume of unwanted spam emails annoys the Internet users; causes spam activities and financial losses. So, spam detection is a serious task to provide a secure electronic environment. Email spam databases usually have multimodal distributions with high overlap, which cause difficulties in separating spam emails from normal emails. Moreover, the number of available labeled emails may be limited. A supervised feature extraction method, which is called cluster space linear discriminant analysis (CSLDA), is proposed in this paper to deal with these difficulties. CSLDA uses the ability of unlabeled testing samples in addition to labeled training ones for estimation of the within-class and between-class scatter matrices. Based on the multimodal distribution of email spam databases, CSLDA clusters the unlabeled testing data for using them in the learning phase of feature extraction. CSLDA uses the testing samples without determination of their labels, and just with obtaining relationship between training and testing samples through clustering. The use of Fisher criterion increases the class discrimination. Moreover, the use of clustered unlabeled samples solves the small sample size problem and provides good performance for multimodal data. The experimental results on spambase dataset indicate the superiority of CSLDA compared to some popular and state-of-the-art feature extraction and spam detection methods, especially in small sample size situations.

Keywords: Classification. Clustering. Discriminant analysis. Email spam.

1. Introduction

The Internet users deal with many threats in the web environment. The web threats may cause loss of private information, identify theft, and financial damage. Electronic mail (Email) as a simple, cheap, and effective tool of communications is used by almost all Internet users. This simple communication tool is attacked by a lot of threats such as spam. Spam emails that are a variety of unwanted or junk mails are Internet mails which are sent to a group of users who have not requested them. The economic cost of spam emails is significantly high. This costly burden consists of the bandwidth, memory, time expended by users, server capacity wasted, and businesses' productivity losses [1], [5]. Spam emails affect millions of users. For instance, according to published reports [6-7], the global ratio of spam in email traffic is more than 70%. In addition, spam emails may cause security breaches. For instance, some major threats presented by spam are phishing [8-9], which bothers users through stealing personal information from the recipient, such as bank account numbers, and spoofing, which is based on misleading [10-11]. So, dealing with spam is a key challenge for information technology research.

Spam detection is a categorization problem where the classes to be predicted are spam and legitimate. Many machine-learning algorithms, such as Bayes classifier [12], [14] and support vector machine [15-16], have been successfully applied to the email spam detection task [17], [19]. Soft machine learning methods such as neural networks need a lot of training samples to learn. In other words, they cannot be trained using small training samples, and therefore, they cannot work using limited training samples [20], [22].

The negative selection algorithm (NSA) inspired by artificial immune system model combined with differential evolution (DE) has been proposed in [2]. In the NSA–DE method, DE generates detectors at the random detector generation phase of NSA, the generated detector distance is maximized and overlapping of detectors is minimized. The spam detection algorithm introduced in [19] uses a wrapper-based feature selection method to extract crucial features. It uses the decision tree classifier and the binary particle swarm optimization (PSO) with mutation operator (MBPSO) as the subset search strategy. In the combined NSA–PSO method [23], PSO is implemented to improve the random detector generation in the NSA algorithm. In [22], the application of the group method of data handling (GMDH) based inductive learning approach is explored in spam detection by automatically identifying content features which effectively distinguish legitimate from spam emails.

Feature extraction methods [24], [27] can be applied for feature transformation of data to increase the class discrimination. Linear discriminant analysis (LDA) [28-29] is a popular and widely used supervised feature extraction method in pattern recognition problems. LDA maximizes the between-class scatter matrix while simultaneously minimizes the within-class scatter matrix. Because of singularity of within-class scatter matrix, LDA fails to work in small sample size situations. The median-mean line based discriminant analysis (MMLDA) method [30] alleviates the negative effect of outliers on the class mean with introducing the median-mean line as an adaptive class-prototype. The within-class scatter matrix is usually singular in small sample size situation. So, to overcome this problem, the principal component analysis can be used first to reduce the dimension of data and then, the MMLDA method is performed in the reduced transformed space. Locality preserving projection (LPP) [31] is an effective manifold learning method, which can be performed supervised or unsupervised. Unsupervised LPP, which uses no class label information, represents the topological structure of data with an adjacency graph, while supervised LPP preserves the local structure of labelled samples by the constructed graph. Unsupervised LPP only considers the relationship between two points when constructing the graph, while supervised LPP only considers samples within the same class during graph construction. The values in similarity matrix are assigned as one when two samples belong to the same class. Otherwise, the values are assigned to be zero.

Email spam datasets usually have multimodal distribution. This characteristic of data degrades the performance of the above mentioned feature extraction methods. To deal with this problem, a supervised feature extraction method is proposed in this paper that uses the

clustering approach to improve the email spam detection performance. The proposed method, which is called cluster space linear discriminant analysis (CSLDA), increases the class discrimination with maximizing the between-class scatters and minimizing the within-class scatters. To deal with the limited availability of labelled training instances, and also to generalize the classifier for classification of non-seen testing samples, CSLDA uses the ability of unlabelled testing samples in addition to labelled training samples. CSLDA clusters the unlabelled testing data and calculates the likelihood that each training sample belong to each of clusters. Therefore, CSLDA obtains the relation between training samples and testing ones through clustering and Bayes theorem to calculate the membership probabilities of training samples in each cluster of testing data. In other words, CSLDA uses the potential of testing samples in the learning phase of feature extraction without determination of their labels. After transformation, the extracted features, which have more class discrimination ability compared to original features, are given to an appropriate classifier for classification. The nearest neighbour classifier is used in this work. The classification results using features extracted by the proposed CSLDA method are compared with features extracted by LDA, MMLDA, supervised LPP, and also with original features for spam base dataset. Moreover, the classification accuracy of CSLDA is compared with NSA–DE [2], MBPSO [19], NSA–PSO [23], and GMDH [22] spam detection methods. The experimental results show better performance of CSLDA compared to other methods, especially while using small training sets.

The remainder of this paper is organized as follows. The CSLDA method is introduced with more details in Section 2. The experimental results are discussed in Section 3. Finally, conclusions are presented in Section 4.

2. Proposed Method

An email spam detection method is introduced in this section that uses a feature transformation before classification. Fig. 1 shows the normalized histogram (probability distribution) of spam and normal data for several features of database, i.e., word_freq_all, word_freq_our, word_freq_order, word_freq_will, word_freq_you, and char_freq_!. Two conclusions can be made from this figure. First, the distribution of email spam data is multimodal. Second, the spam and normal classes have overlap in the most dimensions of feature space. Based on these findings, we propose a feature extraction method in this paper that solves the difficulties of spam datasets. The proposed method, based on the multimodal nature of data, clusters the testing data and uses the ability of clustered unlabeled testing samples to solve the small sample size problem. The proposed method increases the class separability using Fisher criterion. The clustering approach, for estimation of scatter matrices, deals with multi-modal data. Moreover, the proposed method, by using both labeled training samples and unlabeled testing ones in the learning phase of feature extraction, generalizes the ability of classifier for classification of unseen email samples. So, the classification accuracy is significantly increased using limited training samples.

The proposed method, which is called cluster space linear discriminant analysis (CSLDA), utilizes the testing samples without determination of their labels, and only with obtaining the relationship between training samples and testing ones. The use of unlabeled samples in the learning phase of feature extraction process provides some advantages. For instance, it generalizes the ability of classifier for discrimination and classification of unseen emails. On the other hand, gathering of labeled samples is a hard, costly and time consuming task, and so, the available training samples are limited. Because of using the ability of unlabeled samples, CSLDA can work using limited training samples with better classification accuracy compared to other feature extraction methods

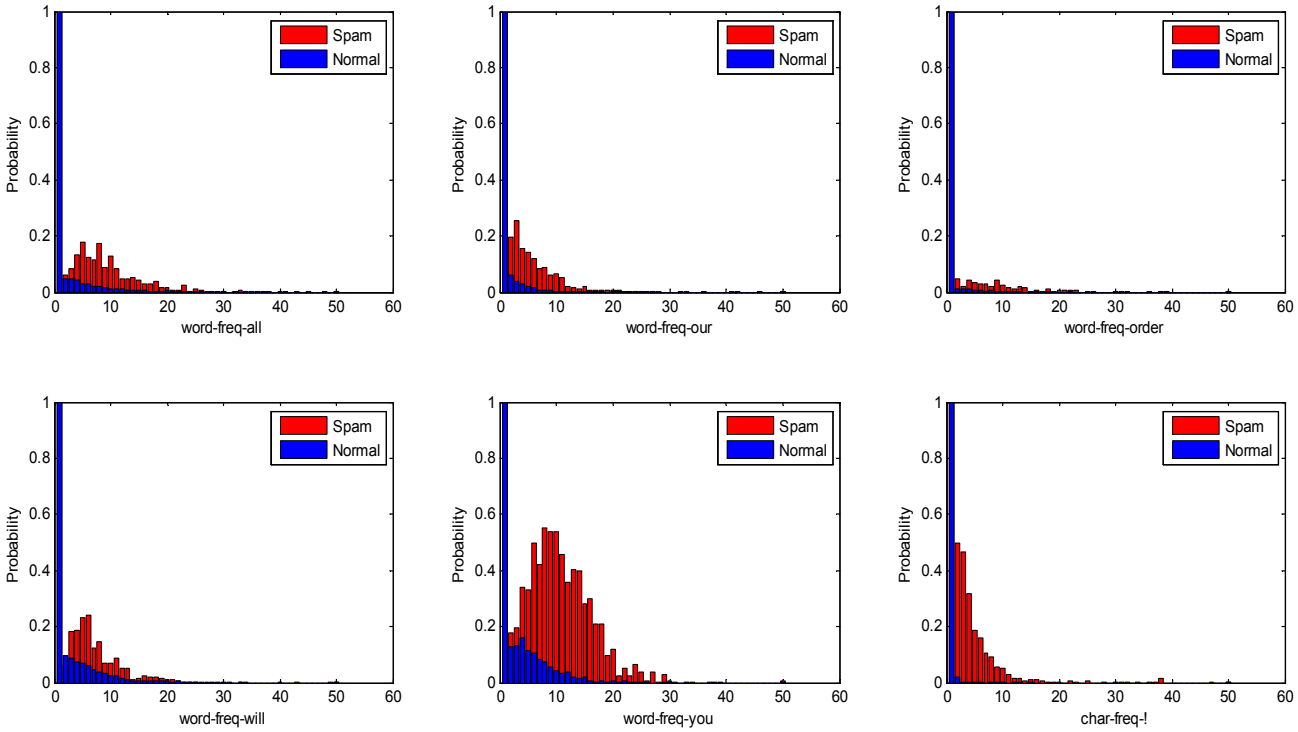


Figure 1. Histogram of email spam and normal data for some email spam features.

The CSLDA method with maximizing the between-class scatters and minimizing the within-class scatters increases the class discrimination, and so improves the classification accuracy. Popular supervised feature extraction methods such as LDA do not work well when there is a small training set. This disadvantage is because of singularity of within-class scatter matrix, and also, non-accurate estimates of both the within-class and between-class scatter matrices. The proposed CSLDA method, by using unlabeled samples in addition to labeled ones, deals with this difficulty. Let $\boldsymbol{x}_{d\times 1}$ be a sample of dataset, which belongs to one of n_c class, where d is the dimensionality, i.e., the number of original features (attributes) of data. Since an information class is insufficiently described by a single well-defined grouping, it is represented by a group of clusters. These clusters are generated from the unlabeled (testing) samples of data. Although clusters can be pure clusters of one information class, it is more probable that each information class is composed of a group of clusters. So, each one of training samples, which has a special class label, can belong to some clusters. Based on this idea, we do a clustering with K clusters on the unlabeled testing data, and then estimate the within-class scatter matrix ($\mathbf{S}_w$) and the between-class scatter matrix ($\mathbf{S}_b$) as follows:

$$\mathbf{S}_w = \sum_{c=1}^{n_c}\sum_{i=1}^{n_{tc}}[(\boldsymbol{x}_{ic}-\boldsymbol{m}_c)(\boldsymbol{x}_{ic}-\boldsymbol{m}_c)^T + \sum_{k=1}^{K}\sum_{\boldsymbol{x}^k\in\boldsymbol{\mathcal{X}}^k} p(k|\boldsymbol{x}_{ic})\,(\boldsymbol{x}^k-\boldsymbol{\mu}_k)(\boldsymbol{x}^k-\boldsymbol{\mu}_k)^T] \quad (1)$$

$$\mathbf{S}_b = \sum_{c=1}^{n_c} n_{tc}(\boldsymbol{M}_c-\boldsymbol{M})(\boldsymbol{M}_c-\boldsymbol{M})^T \quad (2)$$

where $\boldsymbol{x}_{ic}$ is ith training sample of class c, $\boldsymbol{m}_c$ is the mean of training samples of class c, n_{tc} is the number of training samples in class c, $\boldsymbol{x}^k$ is a testing sample that belongs to cluster k, $\boldsymbol{\mathcal{X}}^k$ is subset of testing samples of cluster k, and $\boldsymbol{\mu}_k$ is the mean of testing samples of cluster k. n_c and K are the number of classes and the number of clusters respectively. $p(k|\boldsymbol{x}_{ic})$ is the likelihood that $\boldsymbol{x}_{ic}$ belongs to cluster k. The combined mean of class c ($\boldsymbol{M}_c$) and the total combined mean ($\boldsymbol{M}$), which are used in $\mathbf{S}_b$, are calculated as follows:

$$\boldsymbol{M}_c = \boldsymbol{m}_c + \left(\frac{1}{n_{tc}}\right)\sum_{i=1}^{n_{tc}}\sum_{k=1}^{K} p(k|\boldsymbol{x}_{ic})\,\boldsymbol{\mu}_k \quad (3)$$

$$\boldsymbol{M} = \left(\frac{1}{n_c}\right)\sum_{c=1}^{n_c}\boldsymbol{m}_c + \left(\frac{1}{K}\right)\sum_{k=1}^{K}\boldsymbol{\mu}_k \quad (4)$$

The mean of training samples in class c ($\boldsymbol{m}_c$) and the mean of testing samples in cluster k ($\boldsymbol{\mu}_k$) are calculated as follows:

$$\boldsymbol{m}_c = \frac{1}{n_{tc}}\sum_{i=1}^{n_{tc}}\boldsymbol{x}_{ic} \quad (5)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{\boldsymbol{x}^k\in\boldsymbol{\mathcal{X}}^k}\boldsymbol{x}^k \quad (6)$$

where N_k is the number of unlabeled testing samples of cluster k. Note that Eq. (1), for calculation of $\mathbf{S}_w$, is composed of two terms. The first term uses only the labeled training samples, and the second term uses only the unlabeled testing samples to calculate the within-class scatter matrix. Actually, ith training sample of class c ($\boldsymbol{x}_{ic}$) can belong to several clusters. The likelihood that $\boldsymbol{x}_{ic}$ belongs to each of K available clusters is obtained. Corresponding to these likelihoods, the testing samples of each cluster are contributed in calculation of $\mathbf{S}_w$ according to the second term of (1). Moreover, the combined mean of class c, i.e., $\boldsymbol{M}_c$, is composed of two terms. The first term is

the mean of training samples of class c, and the second term calculates the weighted mean of testing samples in clusters. In other words, each testing sample of cluster k is contributed in the calculation of $\boldsymbol{M}_c$ corresponding to the membership probability of training sample of class c to the cluster k. The total mean $\boldsymbol{M}$ is also composed of two terms where the first term and the second term are the mean of training samples and the mean of testing samples respectively. With assuming the Gaussian distribution for data, the likelihood that sample $\boldsymbol{x}_{ic}$ belongs to cluster k is given by:

$$p(\boldsymbol{x}_{ic}|k) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_k|}} exp\left(-\frac{1}{2}(\boldsymbol{x}_{ic}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{ic}-\boldsymbol{\mu}_k)\right) \quad (7)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and the covariance matrix of cluster k respectively. The matrix $\boldsymbol{\Sigma}_k$ is estimated as follows:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k-1}\sum_{\boldsymbol{x}^k\in\boldsymbol{X}^k}(\boldsymbol{x}^k-\boldsymbol{\mu}_k)\,(\boldsymbol{x}^k-\boldsymbol{\mu}_k)^T \quad (8)$$

We can assume that all clusters have a spherical distribution, and so, their covariance matrices are unit diagonal to avoid the covariance matrix estimation for each cluster. This assumption is reasonable since each cluster has relatively small variance and moreover, we use the K-means algorithm for clustering where it favors the generation of clusters that are hyperspherical. Therefore, the probability distribution function for cluster k can be calculated as:

$$p(\boldsymbol{x}_{ic}|k) = \frac{1}{\sqrt{(2\pi)^d}} exp\left(-\frac{1}{2}dist(\boldsymbol{x}_{ic},\boldsymbol{\mu}_k)\right) \quad (9)$$

where

$$dist(\boldsymbol{x}_{ic},\boldsymbol{\mu}_k) = (\boldsymbol{x}_{ic}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_{ic}-\boldsymbol{\mu}_k) \quad (10)$$

is the Euclidean distance from $\boldsymbol{x}_{ic}$ to $\boldsymbol{\mu}_k$. The likelihood that correct cluster is k for training sample $\boldsymbol{x}_{ic}$ is determined using Bayes' theorem, as follows:

$$p(k|\boldsymbol{x}_{ic}) = \frac{p(\boldsymbol{x}_{ic}|k)p(k)}{\sum_{k=1}^{K}p(\boldsymbol{x}_{ic}|k)p(k)} \quad (11)$$

that $p(k) = \frac{N_k}{N_{test}}$ is the prior probability of cluster k where N_{test} is the total number of testing samples. With combining the equations (9) and (11), we have:

$$p(k|\boldsymbol{x}_{ic}) = \frac{exp\left(-\frac{1}{2}dist(\boldsymbol{x}_{ic},\boldsymbol{\mu}_k)\right)p(k)}{\sum_{k=1}^{K}exp\left(-\frac{1}{2}dist(\boldsymbol{x}_{ic},\boldsymbol{\mu}_k)\right)p(k)} \quad (12)$$

After calculating the scatter matrices, we use the following regularization method to deal with the singularity of $\mathbf{S}_w$ and to increase the classification accuracy:

$$\mathbf{S}_w = 0.5\mathbf{S}_w + 0.5diag(\mathbf{S}_w) \quad (13)$$

To maximize the between-class scatter matrix and to minimize the within-class scatter matrix, the Fisher criterion, $\max tr\left(\boldsymbol{S}_{\mathrm{w}}{}^{-1}\boldsymbol{S}_{\mathrm{b}}\right)$, is used. For extracting p features from d original features, the p eigen vectors of $\boldsymbol{S}_{\mathrm{w}}{}^{-1}\boldsymbol{S}_{\mathrm{b}}$ associated with the largest p eigenvalues of $\boldsymbol{S}_{\mathrm{w}}{}^{-1}\boldsymbol{S}_{\mathrm{b}}$ compose the projection matrix $\boldsymbol{A}$. So, we have: $\boldsymbol{y}_{p\times1} = \boldsymbol{A}_{p\times d}\boldsymbol{x}_{d\times1}$. Finally, the extracted features ($\boldsymbol{y}$) are given to a nearest neighbor classifier for data classification.

3. Experiments

In this section, we evaluate the performance of CSLDA compared to LDA, MMLDA, supervised LPP, and original features for spambase dataset. Moreover, the classification accuracy of CSLDA is compared with NSA–DE [2], MBPSO [19], NSA–PSO [23], and GMDH [22] spam detection methods.

3.1 Dataset and evaluation measures

The spambase dataset [32], acquired from email spam messages, is composed of 4601 emails where 1813 (39%) of them are made to be spam emails and 2788 (61%) of them are identified as normal or non-spam messages. Unlike most corpuses, which are provided in the raw form, acquisition of this corpus is pre-processed. Each sample of this dataset is a 58-dimensional vector where 48 of the features are presented by words that enlisted as the most unlabeled words for the class spam and are generated from the original emails with the absence of a stemming or stop list. The remained 6 features are the percentage of presence of special characters, i.e., ";", "(", "[", "!", "$", and "#". A representation of various measures of presence of capital letters in the text of messages is given by other 3 features. Last feature is the class label and indicates that an instance is spam or non-spam by 1 and 0 respectively. The description of features for this database is given in Table 1. The training samples are chosen randomly from entire datasets and the remained samples are used for testing. Each experiment is repeated 10 times and the average classification results are reported. To evaluate the performance of the proposed method in different sizes of training set, we perform extensive experiments with various training sample sizes. Out of the entire dataset, n% of samples in each class is used for training and the remaining (100-n)% of class samples are used for testing, where n=1, 5, 10, 30, 50, and 75.

Several evaluation measures are used for assessment of email spam detection or any binary classification method [33-34]. These measures consist of confusion matrix, spam classification accuracy (that is equal to true positive rate (TPR) and spam recall (R)), legitimate classification accuracy, false positive rate (FPR), spam precision (P), F-measure, and overall classification accuracy. The confusion matrix shows the actual and predicted classification of each class. Other evaluation measures can be calculated from the confusion matrix as represented in Table 2. In this context, positive and negative refer to email considered as spam and legitimate respectively. The number of spam emails that are correctly detected is denoted by true positive (TP), the number of legitimate emails that are falsely classified as spam is denoted by false positive (FP), the number of spam emails that are falsely classified legitimate is denoted by false negative (FN), and the number of legitimate emails that are correctly classified is denoted by true negative (TN). By considering different thresholds, the tradeoff between TPR and FPR is described visually by the receiver

Table 1. The description of features in spambase dataset.

1	word_freq_make	16	word_freq_free	31	word_freq_telnet	46	word_freq_edu
2	word_freq_address	17	word_freq_business	32	word_freq_857	47	word_freq_table
3	word_freq_all	18	word_freq_email	33	word_freq_data	48	word_freq_conference
4	word_freq_3d	19	word_freq_you	34	word_freq_415	49	char_freq_;
5	word_freq_our	20	word_freq_credit	35	word_freq_85	50	char_freq_(
6	word_freq_over	21	word_freq_your	36	word_freq_technology	51	char_freq_[
7	word_freq_remove	22	word_freq_font	37	word_freq_1999	52	char_freq_!
8	word_freq_internet	23	word_freq_000	38	word_freq_parts	53	char_freq_$
9	word_freq_order	24	word_freq_money	39	word_freq_pm	54	char_freq_#
10	word_freq_mail	25	word_freq_hp	40	word_freq_direct	55	capital_run_length_average
11	word_freq_receive	26	word_freq_hpl	41	word_freq_cs	56	capital_run_length_longest
12	word_freq_will	27	word_freq_george	42	word_freq_meeting	57	capital_run_length_total
13	word_freq_people	28	word_freq_650	43	word_freq_original	58	class_label
14	word_freq_report	29	word_freq_lab	44	word_freq_project		
15	word_freq_addresses	30	word_freq_labs	45	word_freq_re		

Table 2. Evaluation measures for assessment of email spam detection

Evaluation measure	Formula
Spam classification accuracy, True positive rate, spam recall	$Acc_{spam} = TPR = R = \frac{TP}{TP + FN}$
Legitimate classification accuracy	$Acc_{leg} = \frac{TN}{FP + TN}$
False positive rate	$FPR = \frac{FP}{FP + TN}$
Spam precision	$P = \frac{TP}{TP + FP}$
F-measure	$F = \frac{2.P.R}{P + R}$
Overall classification accuracy	$Acc = \frac{TP + TN}{TP + FP + FN + TN}$

operating characteristic (ROC) curve [35]. The area under the curve (AUC) [36] an important scalar measure, which is calculated from the ROC curve. Having a classifier with higher AUC value is desirable. The useful classifiers would have AUC values in the range [0.5,1] and an ideal classifier has AUC value equal to 1.

3.2 Experimental Result

The same training samples are used for obtaining the feature transformation matrix in CSLDA, LDA, MMLDA, and supervised LPP methods. The features extracted by them, and also the original features, are given to a nearest neighbor classifier with Euclidean distance. In CSLDA, the number of clusters is considered equal to the number of classes, i.e., $K = n_c = 2$. The classification results acquired by 1%, 5% 10%, 30%, 50%, and 75% training samples for spambase dataset are reported in Table 3. The best results are shown in bold. The ROC curves for spambase dataset are also shown in Fig. 2. The following conclusions can be found from the results:

1- CSLDA provides the highest classification accuracy in terms of different evaluation measures by using both small and large training sets.

2- Although in some cases, when there is large enough training samples, the LDA or original features provide better results than CSLDA in terms of spam/legitimate classification accuracy (Acc_{spam}/Acc_{leg}), false positive rate (FPR), or spam precision (P); but in all cases, by using small or large training sets, CSLDA achieves the highest classification accuracy in terms of F-measure (F), Overall classification accuracy (Acc), and the area under the ROC curve (AUC), which consider both the spam and legitimate classification accuracies.

3- The superiority of CSLDA compared to other methods is significant by using small training sets. It is expected, because CSLDA uses the ability of unlabeled testing samples in addition to labeled training ones to deal with the small sample size situation.

4- While LDA has low efficiency by using small training sets, because of singularity of within-class scatter matrix, it provides the best classification results, after CSLDA, by using large training sets.

5- Original features provide more classification accuracy compared to features extracted by MMLDA and LPP.

6- The efficiency of LPP is superior to MMLDA.

7- By increasing the number of used training samples, the classification results are improved, and the ROC curves are closed to the upper left corner of the ROC space. Moreover, the efficiencies of all methods become nearly close together.

From the obtained results, it is concluded that the CSLDA method is an efficient feature extraction method that produces features which can significantly improve the performance of email spam detection especially when there is limited training samples. In addition to feature extraction methods, the performance of CSLDA is compared with NSA–DE [2], MBPSO [19], NSA–PSO [23], and GMDH [22] spam detection methods in terms of overall classification accuracy. The results acquired by 75% training samples for spambase dataset are reported in Table 4. The comparison results show the superiority of CSLDA respect to other spam detection frameworks.

Table 3. The classification results for CSLDA compared to LDA, MMLDA, LPP, and original features.

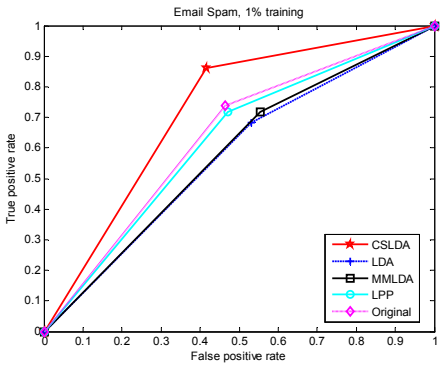
5% training samples							
	$Acc_{spam} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
CSLDA	**80.53**	**83.75**	**19.47**	**86.87**	**85.28**	**82.48**	**82.14**
LDA	55.21	72.17	44.79	71.25	71.70	65.49	63.69
MMLDA	61.67	74.71	38.33	74.98	74.85	69.57	68.19
LPP	61.83	74.21	38.17	74.94	74.57	69.33	68.02
Original	63.71	76.61	36.29	76.45	76.53	71.53	70.16

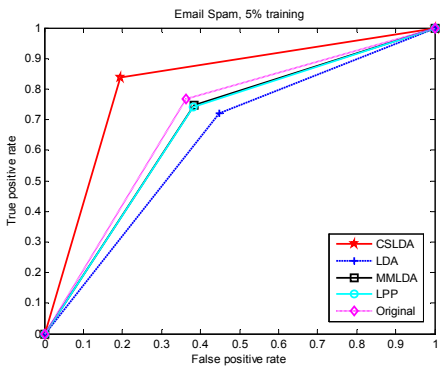
10% training samples							
	$Acc_{spam} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
CSLDA	82.52	**87.30**	17.48	88.48	**87.89**	**85.42**	**84.91**
LDA	**96.80**	23.67	**3.20**	**91.92**	37.65	52.49	60.24
MMLDA	61.89	77.01	38.11	75.65	76.32	71.05	69.45
LPP	64.59	77.47	35.41	77.09	77.28	72.40	71.03
Original	66.08	81.99	33.92	78.80	80.37	75.72	74.04

30% training samples							
	$Acc_{spam} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
CSLDA	**89.02**	**92.79**	**10.98**	**92.86**	**92.82**	**91.31**	**90.91**
LDA	85.60	91.18	14.40	90.69	90.93	88.98	88.39
MMLDA	72.31	81.96	27.69	81.99	81.97	78.16	77.13
LPP	76.28	83.82	23.72	84.46	84.14	80.85	80.05
Original	79.26	87.02	20.74	86.58	86.80	83.96	83.14

50% training samples							
	$Acc_{spam} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
CSLDA	**92.61**	93.87	**7.39**	**95.13**	**94.49**	**93.37**	**93.24**
LDA	91.23	**94.19**	8.77	94.29	94.24	93.02	92.71
MMLDA	83.34	87.37	16.66	88.97	88.17	85.79	85.36
LPP	84.28	88.77	15.72	89.67	89.22	87.00	86.53
Original	87.26	91.14	12.74	91.67	91.40	89.61	89.20

75% training samples							
	$Acc_{spam} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
CSLDA	**96.53**	96.99	**3.47**	**97.72**	**97.35**	**96.81**	**96.76**
LDA	95.31	**97.63**	4.69	96.97	97.30	96.72	96.47
MMLDA	92.94	94.37	7.06	95.36	94.86	93.81	93.65
LPP	93.00	94.58	7.00	95.41	94.99	93.96	93.79
Original	94.21	95.95	5.79	96.22	96.08	95.26	95.08





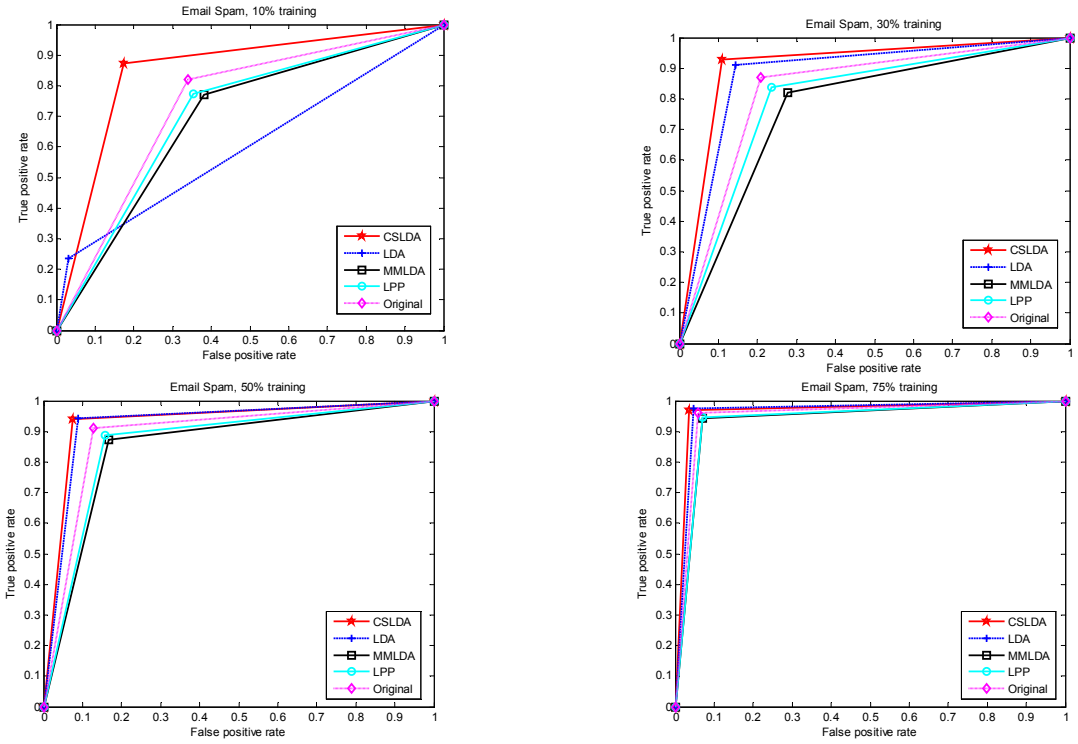


Figure2. ROC curves achieved by 1%, 5% 10%, 30%, 50%, and 75% training samples for spambase dataset

Table 4. The overall classification accuracy of CSLDA compared to other spam detection frameworks

	CSLDA	NSA–DE	MBPSO	NSA–PSO	GMDH
Acc	96.81	80.66	94.27	82.62	91.7

4. Conclusion

An email spam detection algorithm, based on a feature extraction method, is proposed in this paper. The proposed CSLDA method significantly increases the class discrimination and deals with available limited training samples by using clustered unlabeled testing samples. The CSLDA method uses the Bayes' theorem to obtain the membership likelihood of each training sample to each of clusters of testing data. Each unlabeled testing sample contributes in estimation of scatter matrices beside the associated labeled training samples corresponding to the membership probabilities. The CSLDA method provides better classification results in terms of different classification measures. This improvement is significant by using limited training samples. For instance, while original features provide 65.88% overall classification accuracy by using 1% samples as training in spambase dataset, CSLDA achieves 75.16% classification accuracy in the same conditions.

Acknowledgment

This work was supported in part by National Elites Foundation. The authors gratefully acknowledge that organization for its support.

References

[1] F. Gillani, E. Al-Shaer, and B. AsSadhan, "Economic metric to improve spam detectors," *Journal of Network and Computer Applications*, vol. 65, pp. 131–143, 2016.
[2] I. Idris, A. Selamat, and S. Omatu, "Hybrid email spam detection model with negative selection algorithm and differential evolution," *Engineering Applications of Artificial Intelligence*, vol. 28, pp. 97-110, 2014.
[3] J. Goodman, D. Heckerman, and R. Rounthwaite, "Stopping spam," *Scientific American*, vol. 292, no. 4, pp. 42–49, 2005.
[4] Leiba B., "Unwanted Traffic, Finding and Defending against Denial of Service, Spam, and Other Internet Flotsam," IEEE Internet Computing, vol. 13, no. 6, pp. 10-13, 2009.
[5] O. Fonseca, E. Fazzion, I. Cunha, P. Las-Casas, D. Guedes, Jr. Meira W, C. Hoepers, K. Steding-Jessen, and M. H. P. Chaves, "Measuring, Characterizing, and Avoiding Spam Traffic Costs," *IEEE Internet Computing*, vol. 20, no. 4, pp. 16-24, 2016.
[6] S. Dinh, T. Azeb, F. Fortin, D. Mouheb, and M. Debbabi, "Spam campaign detection, analysis, and investigation," *Digital Investigation*, vol. 12, pp. S12-S21, 2015.
[7] C. Laorden, X. Ugarte-Pedrero, I. Santos, B. Sanz, J. Nieves, and P. G. Bringas, "Study on the effectiveness of anomaly detection for spam filtering," *Information Sciences*, vol. 277, pp. 421–444, 2014.
[8] T-C. Chen, T. Stepan, S. Dick, and J. Miller, "An investigation of implicit features in compression-based

learning for comparing webpages," *Pattern Analysis and Applications*, vol. 19, pp. 397–410, 2016.
[9] G. A. Montazer, and S. ArabYarmohammadi, "Detection of Phishing Attacks in Iranian E-banking Using a Fuzzy Rough Hybrid System," *Applied Soft Computing*, vol. 35, pp. 482–492, 2015.
[10] S. Heron, "Technologies for spam detection," *Netw. Secur*, vol. 1, pp. 11–15, 2009.
[11] A. Hamdan Mohammad, and R. Abu Zitar, "Application of genetic optimized artificial immune system and neural networks in spam detection," *Appl. Soft Comput*. , vol. 11, pp. 3827–3845, 2011.
[12] T. Oda, and T. White, "Immunity from Spam: an analysis of an artificial immune system for junk email detection. Jacob, C., Pilat, M., Bentley, P., Timmis, J. (Eds.)," *Artificial Immune Systems*. Springer, Berlin, Heidelberg, vol. 3627, pp. 276–289, 2005.
[13] D. H. Shih, H. S. Chiang, and C. D. Yen, "Classification methods in the detection of new malicious emails," *Information Sciences*, vol. 172, pp. 241–261, 2005.
[14] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem," *Decision Support Systems*, vol. 46, no.4, pp. 853–864, 2009.
[15] R. Guangchen, and T. Ying, "Intelligent detection approaches for spam," *Third International Conference on Natural Computation*, ICNC2007, pp. 672–676, 2007.
[16] A. Kolcz, and J. Alspector, "SVM-based filtering of e-mail spam with content specific misclassification costs," *Proc. of the TextDM'01 workshop on text mining-held at the 2001 IEEE international conference on date mining*, 2001.
[17] M. Wamli, T. Dat, and S. Dharmendra, "A novel spam email detection system based on negative selection," *Proc. of the Fourth International Conference on Computer Science and Convergence Information Technology*, 2009.
[18] E.-S. M. El-Alfy, and A. A. AlHasan, "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm," *Future Generation Computer Systems*, vol. 64, pp. 98–107, 2016.
[19] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowledge-Based Systems*, vol. 64, pp. 22-31, 2014.
[20] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999.
[21] M-C. Su, H-H. Lo, and F-H. Hsu, "A neural tree and its application to spam e-mail detection," *Expert Systems with Applications*, vol. 37, pp. 7976–7985, 2010.
[22] E-S. El-Alfy, and R. Abdel-Aal, "Using GMDH-based networks for improved spam detection and email feature analysis," *Applied Soft Computing*, vol. 11, no. 1, pp. 477-488, 2011.
[23] I. Idris, A. Selamat, N. Thanh Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system," *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 33-44, 2015.
[24] M. Imani, and H. Ghassemian, "High-Dimensional Image Data Feature Extraction by Double Discriminant Embedding," *Pattern Analysis and Applications*, vol. 20, no. 2, pp. 473–484, 2017.
[25] C. Vicient, D. Sánchez, and A. Moreno, "An automatic approach for ontology-based feature extraction from heterogeneous textualresources," *Eng. Appl. Artif. Intell*. , vol. 26, no. 3, pp. 1092-1106, 2013.
[26] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, "Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification," *Computer Networks*, vol. 53, pp. 835–848, 2009.
[27] M. Imani, and H. Ghassemian, "Binary coding based feature extraction in remote sensing high dimensional data," *Information Sciences*, vol. 342, pp. 191–208, 2016.
[28] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* San Diego: Academic Press Inc, 1990.
[29] J-G. Wang, E. Sung, and W-Y. Yau, "Incremental two-dimensional linear discriminant analysis with applications to face recognition," *Journal of Network and Computer Applications*, vol. 33, pp. 314–322, 2010.
[30] J. Xu, J. Yang, Z. Gu, and N. Zhang, "Median–mean line based discriminant analysis," *Neurocomputing*, vol. 123, pp. 233–246, 2014.
[31] X. F. He, and P. Niyogi, Locality preserving projections. In Proc. Adv. *Neural Inf. Process. Syst*., vol. 16, pp. 153–160, 2004.
[32] Hopkins M, Reeber E, Forman G, and Suermondt J., "Spam Base Dataset," *Hewlett-Packard Labs* 1999.
[33] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press 2008.
[34] L. Araujo, and J. Martinez-Romo, "Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 581-590, 2010.
[35] Fawcett T., "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.
[36] J.A. Hanley, and B.J.A. McNeil, "Method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

Maryam Imani received the B.Sc. and M.Sc. degrees in electrical engineering from Shahed University, Tehran, Iran, and the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran in 2009, 2011, and 2015 respectively. She continued her research in Tarbiat Modares University as a postdoc. She is an Assistant professor of the Faculty of Electrical and Computer Engineering at Tarbiat Modares University, Tehran, Iran. Her research interests include pattern recognition, signal and image processing, information analysis, and remote sensing.

Email: maryam.imani@modares.ac.ir

Gholam Ali Montazer received his B.Sc. degree in Electrical Engineering from Kh.N. Toosi University of Technology, Tehran, Iran, in 1991, his M.Sc. degree in Electrical Engineering from Tarbiat Modares University, Tehran, Iran, in 1994, and his Ph.D. degree in Electrical Engineering from the same university, in 1998. He is an Associate Professor of the Department of Information Engineering at Tarbiat Modares University, Tehran, Iran. His areas of research include Information Engineering, Knowledge Discovery, Intelligent systems, E-Learning and Image Mining.

Email: montazer@modares.ac.ir

Paper Handling Data:

Submitted: 28.04.2017
Received in revised form: 07.05.2018
Accepted: 02.06.2018
Corresponding author: Dr. Maryam Imani
Faculty of Electrical and Computer Engineering, Tarbiat Modares University