The ISC Int'l Journal of Information Security

ISeCure

July 2017, Volume 9, Number 2 (pp. 147–159)

http://www.isecure-journal.org

Phishing Website Detection Using Weighted Feature Line Embedding

Maryam Imani [1,*], and 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*

ARTICLE INFO.

Article history:

Received: 28 April 2017

Revised: 23 July 2017

Accepted: 30 July 2017

Published Online: 30 July 2017

Keywords:

Phishing Detection, Feature Extraction, Feature Line, Virtual Training.

ABSTRACT

The aim of phishing is tracing the users' s private information without their permission by designing a new website which mimics the trusted website. The specialists of information technology do not agree on a unique definition for the discriminative features that characterizes the phishing websites. Therefore, the number of reliable training samples in phishing detection problems is limited. Moreover, among the available training samples, there are abnormal samples that cause classification error. For instance, it is possible that there are phishing samples with similar features to legitimate ones and vice versa. A supervised feature extraction method, called weighted feature line embedding, is proposed in this paper to solve these problems. The proposed method virtually generates training samples by utilizing the feature line metric. Hence, it can solve the small sample size problem. Moreover, by assigning appropriate weights to each pair of feature points, it corrects the undesirable quality of abnormal samples. The features extracted by our method improve the performance of phishing website detection specially by using small training sets.

1 Introduction

By emerging information technology, many consumers use the online channels to deal with many products and services by using a personal computer. Therefore, they have multiple online accounts such as bank accounts, email accounts, and social network accounts. This technological trend causes a rising threat of online identity theft. In addition to individual users, Internet is also important for organizations doing online trading. But, Internet security for commercial transactions is not so confident. Phishing is one of the main forms of web threats where attackers impersonate website of an honest organization aiming to acquire the private information of Internet users such as passwords and social security numbers [1, 2]. Phishing websites are created by dishonest persons to defraud the honest users. Typically a phishing attack starts by sending an email which seems to be from an honest enterprise to victims urging them to validate or update their information by following a URL link within the email spam [3]. These websites are visually similar to genuine ones. There are a variety of cues for phishing website recognition. These cues are seen in URLs (e.g., https, number of slashes), hyperlinks (e.g., number of in/out links), image pixels (e.g., pixel colors), and features extracted from the textual content of websites (e.g., lexical measures, spelling)

* Corresponding author.

Email addresses: maryam.imani@modares.ac.ir (M. Imani), montazer@modares.ac.ir (G. Montazer)

ISSN: 2008-2045

[4]. For example, two instances of phishing websites and cues to recognize them are shown in Figure 1 [5]. The received phishing reports is having a considerable growth in the recent years. Figure 2 compares the total number of phishing reports received in different months of 2005, 2010, and 2015 [6]. To combat phishing, there are many ways such as legal solutions, educational solutions, and technical solutions [7]. In the legal solution, phishing is added to the computer crime list and phishers are arrested. In the educational solution, the Internet users are learnt to inspect the security indicators within the website. In the technical solution, the academic studies are used to offer commercial and non-commercial anti-phishing solutions. The first and second solutions have some weaknesses. For instance, it is difficult to trace phishers and also Internet users require a long time to learn phishing methods. So, the anti-phishing solution is preferred. However, most of anti-phishing solutions are unable to make perfect decisions on whether website is phishy or legitimate. There are two main approaches to design technical anti-phishing solutions. 1- The use of blacklist approach that compares the requested URL with the predefined phishing URLs. This solution is not very desirable because, the blacklist usually cannot cover all phishing websites. 2- The feature-based methods which collect several features from the website to classify it as either phishy or legitimate. In other words, these methods pick a set of discriminate features to distinguish the type of websites [8, 9]. Our focus in this work is on the feature based methods.

Many machine learning approaches and data mining techniques are suggested for phishing detection problems [10–15]. Soft machine learning methods such as neural networks need a lot of training samples in the learning phase [16, 17]. In other words, they cannot learn using small training samples, and therefore, they cannot work using limited training ones. Different types of feature extraction approaches are introduced to increase the class discrimination in varied ill-posed pattern recognition problems [18]. Several metric learning methods are proposed for feature extraction [19–21]. A robust distance metric whose aim is learning a feature subspace in which the sample points in the same class are as near as possible while the ones in different classes are as far as possible is utilized in [22]. The kernelized versions of metric learning methods are suggested to deal with disadvantages of considering just liner subspaces [23]. In [24], in addition to interclass separability, the label consistency is also considered in the learning process of the discriminative distance metric. Linear discriminant analysis (LDA) is the most popular supervised feature extraction method for optimizing the Fisher criterion [25, 26]. By maximizing the Fisher criterion, the between-class scatters are maximized while simultaneously the within-class scatters are minimized. Thus, the best discriminant vectors can be derived. LDA needs a large training set to estimate the scatter matrices. So, LDA fails to work in small sample size situation because of singularity of within-class scatter matrix. Local preserving projection (LPP) is the another popular and effective feature extraction method used to maintain the locality of the samples structure [27]. LPP can be implemented either supervised or unsupervised. LPP minimizes the distance between neighbouring points and preserves the local structure of samples by the constructed graph. Unsupervised LPP represents the topological structure of samples without using class label information and works well for sample construction but not for classification. On the other hand, supervised LPP only considers samples within the same class during graph construction. In other words, when two samples belong to the same class, the values in the similarity matrix are defined as one; otherwise the values are assigned to be zero. LPP has no need to estimate the statistics of data, and so is efficient in small sample size situations. Unlike the construction of a point-based relationship in the eigenspace projection approaches such as LPP, the nearest feature line embedding (NFLE) method uses a line-based relationship [28]. In other words, NFLE embeds the distance measurement of a feature line through discriminant analysis. Methods such as LPP construct an adjacency matrix to represent the point to point connectivity relationship between neighbouring points. But, since a small number of prototype samples are provided during the training phase, the relationship is poorly modelled, and also, many non-prototype samples under varied conditions are not available during training. But, in NFLE, the linear combinations of original labelled samples virtually generate the non-prototype samples. This point-to-line measurement in NFLE achieves better classification results compared to point-to-point measurement in LPP. Because of the following reasons, gathering of reliable training samples in phishing website detection applications is a hard task:

(1) Because of high similarity between legitimate and phishing websites, the collection of a data that covers all possible discriminating features is difficult. The information technology researchers do not agree on a definition of discriminative features separating real websites from the fake ones. So, usually a small training set is available.
(2) Among the available training samples, there are unreliable samples which have doubtful labels. The presence of these abnormal training sam-

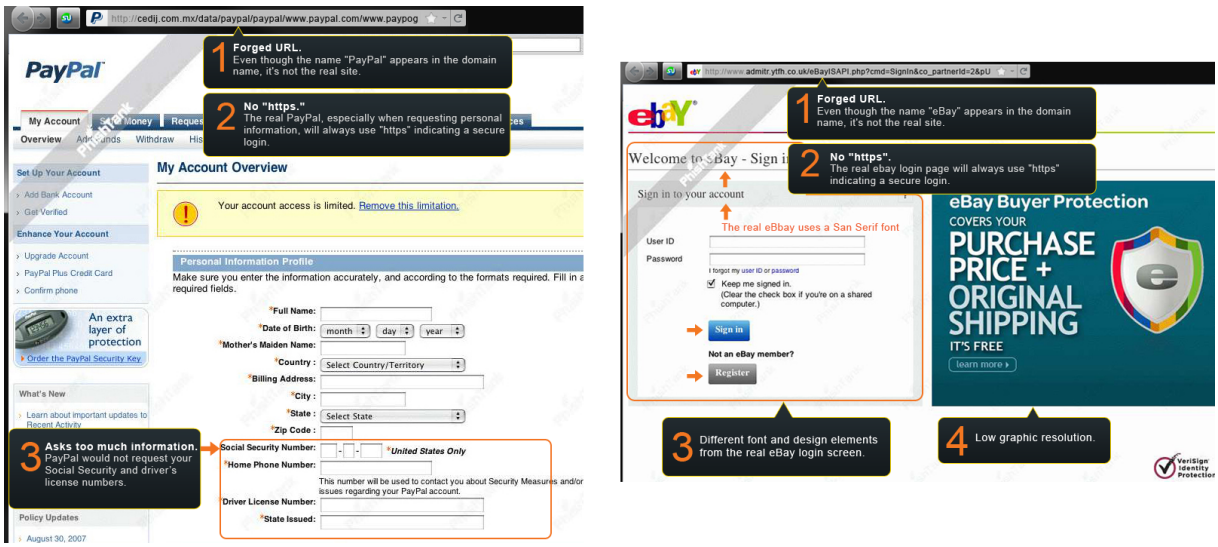

Figure 1. Two phishing websites and cues to recognize them [5].

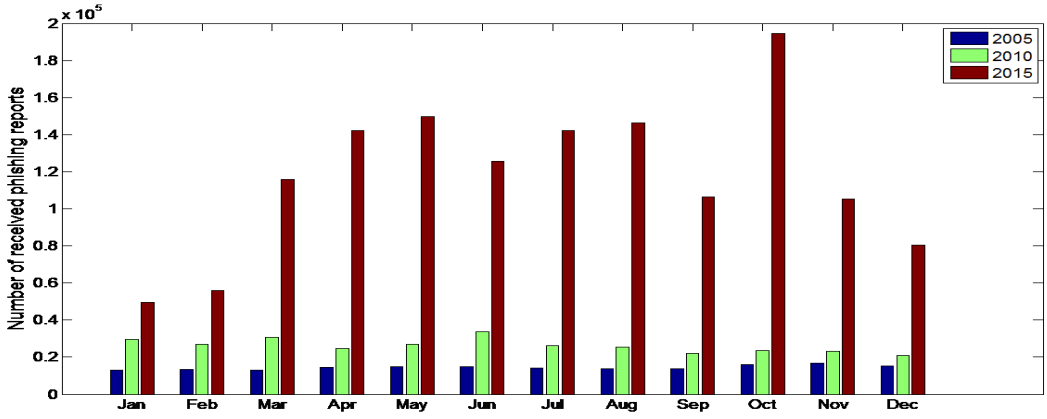

Figure 2. The growth of received phishing reports in different months of 2005, 2010, and 2015 [6].

ples can degrade the probability of correct detection. From one hand, there are some samples belonging to the legitimate class which have features similar to the phishing ones or vice versa. From the other hand, there are within-class legitimate samples that are not similar together or the same class phishing samples that are far from each other in the feature space.

An improved version of NFLE is proposed in this paper. The proposed method, which is called weighted feature line embedding (WFLE), similar to NFLE, embeds the FL metric in the discriminant analysis to produce virtual samples and enlarge the training set. So, it deals with the small sample size problem. In contrast to NFLE, WFLE estimates the scatter matrices in a weighted form where the appropriate weights that are introduced to correct the undesirable quality of unreliable training samples. In this work, we consider training samples in two general groups: 1- normal samples (similar samples belonging to the same class or non-similar samples belonging to the different classes) which are desirable and appropriate for classification, and 2- abnormal samples (non-similar samples belonging to the same class or similar samples belonging to the different classes) which are undesirable and cause classification error. WFLE by considering the appropriate weights corresponding to each pair of the feature points in estimation of scatter matrices degrades the negative effects of abnormal training samples. So, it decreases the classification error. The features extracted by WFLE are given to a nearest neighbour (NN) classifier to find the label of testing samples. The efficiency of WFLE, in phishing detection problem, is compared to LDA, supervised LPP, NFLE, and original features. The experimental results show the superior performance of WFLE especially when there is a limited number of training samples. Some contributions of this work are as follows: 1-The phishing website detection in small sample size situation is assessed. Most of the supervised phishing website detection methods, such as neural networks,

fail to work using limited training samples. 2- The proposed WFLE method, by producing virtual training samples, not only deals with small sample size problem, but also infers the intra-class variations of data through interpolating and extrapolating each pair of available original training samples by using the feature line metric. 3- WFLE rectifies the disadvantages of NFLE by removing the initial feature reduction through the principal component analysis (PCA) transformation, and also by adding correcting weights for estimation of scatter matrices. It decreases the classification error due to the negative effects of abnormal training samples. The reminder of this paper is organized as follows. The related works are reviewed in Section 2. The proposed WFLE method is introduced in Section 3. The used dataset and the assessment measures are presented in Section 4. Moreover, experimental results are discussed this section. Finally Section 5 concludes the paper.

2 Related Work

Let $\{x_i\}_{i=1}^N; x_i \in \mathbb{R}^d$ denote the data in the original feature space where N is the number of samples and d is the dimensionality of data. The low dimensional representation of data in the projected feature space obtained by a feature transformation method is denoted by $\{y_i\}_{i=1}^N; y_i \in \mathbb{R}^m$ where m is the number of extracted features, i.e., the dimensionality of the projected subspace. It is assumed that there is a mapping function $f : \mathbb{R}^d \to \mathbb{R}^m$ which can map every original sample x_i to $y_i = f(x_i)$ such that the most information of the original feature space is kept in a lower dimensional projected subspace. This mapping is usually considered as a $d \times m$ projection matrix A:

$$y_i = f(x_i) = A^T x_i$$

The projection feature matrix A can be obtained supervised where the labelled samples are used for training or unsupervised where no labelled samples are used. Unsupervised methods usually work well for dimensionality reduction or sample reconstruction but not for classification.

2.1 LPP

LPP is a popular linear feature extraction method which preserves the neighbourhood structure of data by building a graph incorporating the neighbourhood information of dataset. Let $X = [x_1, x_2, \ldots, x_n]$ be the available data. The aim is to provide the projected data $Y = A^T X$ where $Y = [y_1, y_2, \ldots, y_n]$; $y_i = A^T x_i$. Assume that G is a graph with n nodes. There is an edge between node i and j if x_i and x_j are close together. There are two ways to find the closeness of samples: ε-neighbourhood and k-nearest neighbours. In the first method, an edge connects nodes i and j if $\|x_i - x_j\| < \varepsilon$ where $\|\cdot\|$ is the usual Euclidean norm. In the second method, an edge connects two nodes i and j if x_i is among k nearest neighbours of x_j or vice versa. The weights of edges, i.e., $w_{ij} (i = 1, \ldots, n; j = 1, \ldots, n)$ compose the sparse symmetric weight matrix W. The weight jointing nodes i and j is zero, $w_{ij} = 0$, if there is no edge between them. A heat kernel such as $w_{ij} = exp\left(\frac{-\|x_i - x_j\|^2}{t}\right)$, or in a simple way, $w_{ij} = 1$ is considered if two vertices i and j are connected together through an edge. Then, the following generalized eigenvector problem is solved by computing the eigenvectors and eigenvalues:

$$XLX^T a = \lambda XDX^T a \tag{1}$$

where D is a diagonal matrix with $D_{ii} = \sum_j w_{ij}$ and $L = D - W$ is the Laplacian matrix. $a_1, a_2, \ldots, a_m$ are the columns of the projection matrix A. They are the solutions of Equation (1) ordered according to the eigenvalues $\lambda_1 < \ldots < \lambda_m$.

Note that LPP can be implemented supervised or unsupervised. In the supervised LPP, only within-class samples are considered during the graph construction. In other words, the similarities between different classes are not assessed and only the similarities between samples within the same class are measured. When two samples x_i and x_j belong to the same class, w_{ij} is defined either as one or a heat kernel (Gaussian function). If not, the zero value is assigned to it.

2.2 LDA

LDA is the best known and widely used supervised feature extraction method which seeks a new feature space of data by maximizing the ratio of the between-class scatters to the within-class scatters. To this end, the LDA projection matrix $A = [a_1, a_2, \ldots, a_m]$ is optimized as follows:

$$a = \arg\max_a \frac{a^T S_b a}{a^T S_w a}$$

where

$$S_b = \sum_{k=1}^{c} n_k (m_k - m)(m_k - m)^T$$

$$S_w = \sum_{k=1}^{c} (\sum_{i=1}^{n_k} (x_{i,k} - m_k)(x_{i,k} - m_k)^T)$$

In the above equations, S_b and S_w are the between-class and within-class scatter matrices, respectively where c denotes the number of classes, n_k is the number of training samples in kth class, m_k is the mean vector of class k, m is the total mean of training samples, and $x_{i,k}$ is ith training sample of class k. The LDA feature extraction method seeks projection

directions on which the data samples within the same class become as close as possible while separating all the data samples from different classes as far as possible.

2.3 NFLE

The nearest feature line (NFL) classifier was first introduced to extend the capability of NN classifier [29]. The NN method just considers the distance from the query feature point to each labelled feature point of a class and chooses the minimum of the obtained distances as the query to class distance. The NFL classifier generalizes the ability of NN due to the feature lines accounting for new conditions which are not represented by the original samples. The NFL method assumes that multiple (more than one) training samples are available per class. By considering a pair of training samples belonging to the same class, the line passing through two samples is called feature line (FL). The NFL method uses the FL metric to obtain intra-class variations by interpolating and extrapolating the template samples. Therefore, FL may approximate variations of class beyond the template samples and generalize the representational ability of two training samples. The NFL classification rule is based on nearest feature line distance, i.e., the distance between the query sample and its projection onto the FL generated from two feature points.

Many methods have used the FL measure in the classification phase in the pattern recognition problems. But, there are only a few works which use the FL metric in the feature extraction phase. A NFL based subspace learning method that adopts FL in both the feature transformation and classification phases is proposed in [30]. In [30], just the within-class scatters, embedded with the FL metric, are calculated rather than the between-class scatters for feature transformation. The method proposed in [31] also embeds the FL metric into the transformation for feature extraction. The method introduced in [31] instead of estimation of scatter matrices and minimizing the within-class scatters and maximizing the between-class scatters, minimizes the mean square distances between the training samples and their corresponding projection points onto the FLs in the transformed feature space. The NFLE method introduced in [28] uses the FL metric for estimation of both within-class and between-class scatter matrices for feature extraction. The FL based methods effectively improve the classification accuracy especially when the number of training samples per class is limited. The small sample size problem frequently encounters in many applications such as face recognition problems and hyperspectral image classification in remote sensing problems [32, 33].

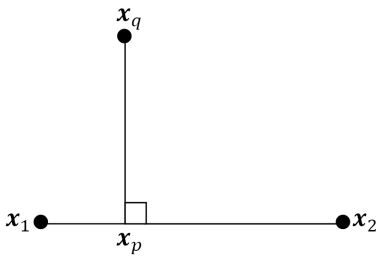

Figure 3. An illustration of feature line.

Let $X = [x_1, x_2, \ldots, x_n]$ to be the training set where $x_i \in R^d$ and n is the number of total training samples. The class label of sample x_i is denoted by $l_{x_i} \in \{1, 2, \ldots, n_c\}$ where $i = 1, 2, \ldots, n$ and n_c is the number of classes. Let n_{tc} be the number of training samples belonging to class c, i.e., $n = \sum_{c=1}^{n_c} n_{tc}$. Assume x_1 and x_2 be two training samples. A FL is a straight line passing through these two samples denoted by $\overline{x_1 x_2}$ (see Figure 3).

The FL approximates linear variants of the class derived from two samples. It provides a virtual training sample of the same class of two training samples x_1 and x_2. The Euclidean distance between the query sample x_q and its projection on $\overline{x_1 x_2}$, i.e., x_p, is called the FL distance, $\|x_q - x_p\|$. The value of the projection point x_p can be calculated as follows:

$$x_p = x_1 + t(x_2 - x_1)$$

where t is the position parameter. Since $x_1\bar{x}_2$ is perpendicular to $x_p\bar{x}_q$, i.e., $(x_p - x_q).(x_2 - x_1) = 0$, we have:

$$t = \frac{(x_q - x_1)^T (x_2 - x_1)}{(x_2 - x_1)^T (x_2 - x_1)}$$

The line passing through two samples x_1 and x_2, not only provides the variations between x_1 and x_2, but also, produces the virtual training sample x_p. Related to the value of parameter t, the interpolation or extrapolation is done:

$$\begin{gathered} t = 1 \rightarrow x_p = x_2 \\ t = 0 \rightarrow x_p = x_1 \\ 0 < t < 1 \rightarrow x_p = \text{linear interpolation}(x_1, x_2) \\ t < 0 \text{ or } t > 1 \rightarrow x_p = \text{linear extrapolation}(x_1, x_2) \end{gathered}$$

In other words, when $0 < t < 1$, x_p is a linear interpolation of x_1 and x_2, and when $t < 0$ or $t > 1$, x_p is a linear extrapolation of x_1 and x_2 (see Figure 4).

The FL measurement can be used into the discriminant analysis for estimation of scatter matrices. The produced virtual samples enlarge the training set and so, the scatter matrices can be accurately estimated. Also the singularity problem of the within-class scatter matrix can be solved by the enlarged training set. Moreover, the useful information of within-class and between-class variations, provided by produced vir-

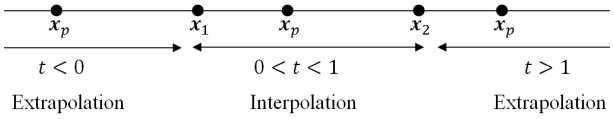

Figure 4. Interpolation and extrapolation related to the value of position parameter.

tual feature points, may increase the class discrimination, and so, improve the classification performance. A number of $\binom{n_{tc}}{2}$ FLs can be produced in class c. The FL metric can be embedded into the discriminant analysis for estimation of within-class scatter matrix (S_w) and between-class scatter matrix (S_b) as follows:

$$S_w = \sum_{i=1}^{n} \sum_{j \in P_i} (x_i - x_{ij})(x_i - x_{ij})^T$$

$$S_b = \sum_{i=1}^{n} \sum_{k \in R_i} (x_i - x_{ik})(x_i - x_{ik})^T$$

where

$$P_i = \{j \mid L_{x_i} = L_{x_{ij}}\}$$
$$R_i = \{k \mid L_{x_i} \neq L_{x_{ik}}\}$$

and x_{ij} and x_{ik} denote the projection of x_i onto the FLs generated by samples belonging to the same class of x_i and samples belonging to different classes of sample x_i. To transform the feature space of data in a linear feature extraction method, a projection matrix A is needed that maps x_i to a new feature space by $(y_i)_{m\times 1} = (A^T)_{m\times d}(x_i)_{d\times 1}$ where d is the dimensionality of original feature vector and m is the number of extracted features. To obtain the projection matrix A using the discriminant analysis, the between-class scatters are maximized while simultaneously the within-class scatters are minimized. To extract m features from d original features, the eigenvectors of $S_w^{-1}S_b$ corresponding to the m largest eigenvalues, composite the projection matrix $A_{d\times m} = [a_1, a_2, \ldots, a_m]$.

3 The proposed WFLE Method

The phishing detection is considered as a two-class classification method where the phishing websites must be separated from the legitimate websites. The proposed method in this work embeds the FL metric in the discriminant analysis in a weighted manner for estimation of scatter matrices S_w and S_b. According to the nature of training samples, an appropriate weight is assigned to the projection of each sample onto the available FLs. In other words, the weight that is assigned to each virtual training sample is related to the nature of training samples composing it. In this work, the original training samples are divided into four categories:

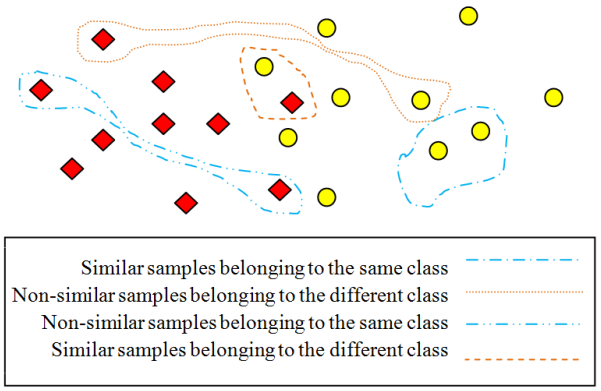

Figure 5. Four categories of normal and abnormal samples in a two-dimensional feature space.

(1) Similar samples belonging to the same class (two samples from the phishing class which have similar features or two samples from the legitimate class which have similar features).
(2) Non-similar samples belonging to the different classes (a sample from the phishing class and a sample from the legitimate class which have non-similar features).
(3) Non-similar samples belonging to the same class (two samples from the phishing class which have non-similar features or two samples from the legitimate class which have non-similar features).
(4) Similar samples belonging to the different classes (a sample from the phishing class and a sample from the legitimate class which have similar features).

The similar samples are defined as samples that there is small distance between them and different samples are samples that there is large distance between them. Euclidean distance is used in this work. From four above introduced categories, the first two categories are normal and have appropriate quality that is desirable for classification. But, the two later categories are abnormal and have inappropriate quality that are undesirable for classification. The abnormal training samples cause error in assigning labels to testing samples, and so, decrease the classification accuracy. The instances of four introduced categories for two classes in a two-dimensional feature space are shown in Figure 5. In the proposed WFLE method, the weights are introduced for correction of undesirable quality of abnormal training samples in the transformed feature space. The between-class and within-class scatter matrices in the WFLE method are defined follows:

$$S_w = \sum_{i=1}^{n} \sum_{j \in P_i} w_{ij}(x_i - x_{ij})(x_i - x_{ij})^T$$

$$S_b = \sum_{i=1}^{n} \sum_{k \in R_i} w_{ik}(x_i - x_{ik})(x_i - x_{ik})^T$$

where w_{ij} and w_{ik} are weights for correction of undesirable quality of abnormal samples in within-class and between-class scatter matrices, respectively. Let p_s and q_s be two samples belonging to the same class of x_i. Then, x_{ij} is the projection of x_i on the FL generated from p_s and q_s. If p_s and q_s be different, i.e., the Euclidean distance between them, $dist(p_s, q_s)$, be large, then, p_s and q_s are considered as abnormal (undesirable) training samples and so, a large weight should be assigned to correct the quality of them. If p_s and q_s be similar together, i.e., the Euclidean distance between them, $dist(p_s, q_s)$, be small, then, p_s and q_s are considered as normal (desirable) training samples and so, a low weight should assign to them in estimation of scatter matrices. Thus, there is a straightforward relationship among the distance between samples and the weighting coefficient in the within-class scatter matrix. So, the weights $w_{ij}(i = 1, 2, \ldots, n; j \in P_i)$ can be calculated as follows:

$$w_{ij} = dist(p_s, q_s) = (p_s - q_s)^T(p_s - q_s)$$

Given p_d and q_d as two samples belonging to the different class of x_i. The projection of x_i on the FL generated from p_d and q_d is denoted by x_{ik}. If p_d and q_d be similar together, i.e., $dist(p_d, q_d)$ be small, then these points are considered as abnormal (undesirable) samples and a larger weight should be assigned for correction of them. In contrast, if p_d and q_d be different respect together, i.e., $dist(p_d, q_d)$ be large, they are normal (desirable) samples and a less weight is needed for correction of them. So, there is an inverse relationship among the distance between samples and the weighting coefficients $w_{ik}(i = 1, 2, \ldots, n; k \in R_i)$ in the between-class scatter matrix:

$$w_{ik} = \frac{1}{dist(p_d, q_d)} = [(p_d - q_d)^T(p_d - q_d)]^{-1}$$

The sample x_i and its projection on the within-class and between-class FLs and their corresponding weights are shown in Figure 6. In addition to defining and using new weighting coefficients for correction of undesirable samples, WFLE has other advantage with respect to NFLE. The NFLE method, at first, i.e., before estimation of scatter matrices, does an initial feature reduction using PCA to deal with singularity of within-class scatter matrix. In the PCA transformation, directions with small variances, which may have considerable class discrimination information, may be lost. So, in the WFLE method, instead of PCA,

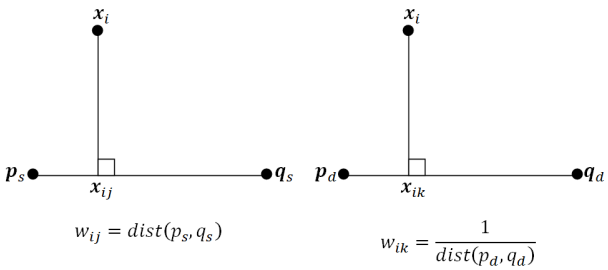

Figure 6. The correction weights in the within-class and between-class FLs.

the following regularization method is used to cope with the singularity of within-class scatter matrix:

$$S_w = \alpha S_w + (1 - \alpha) diag(S_w)$$

where $0 \leq \alpha \leq 1$ is a free parameters. According to the experiments, the final detection result has no significant sensitivity to the value of parameter α. The experiments show that $\alpha = 0.5$ is an appropriate value which can provide good classification or detection results for different types of datasets. So, the same as what is proposed in [34], $\alpha = 0.5$ is chosen in this paper. By maximizing $tr(S_w^{-1} S_b)$, the projection matrix of WFLE is obtained. If training samples have desirable quality (similar samples belonging to the same class or non-similar samples belonging to different classes), they are appropriate for classification and they do not need correction in the transformed feature space. But, if training samples have undesirable quality (non-similar samples belonging to the same class or similar samples belonging to different classes), they are inappropriate for classification, and cause error in the classification result. Thus, more weight should be assigned to undesirable samples for improving the class discrimination. Finally, the features extracted by WFLE, are given to a NN classifier for data classification.

The proposed WFLE method has the following novelties and advantages: 1- It maximizes the separability between legitimate and phishing samples in the feature space. 2- In contrast to feature extraction methods such as LPP and LDA, which represent the point to point connectivity relationship between neighbouring samples, WFLE represents the point to line measure by virtually generation of non-prototype samples through the feature line metric. The virtual non-prototype samples are not available under various conditions among the original training samples, and so, the use of them can provide better detection results. 3- The virtual training samples achieved by feature line metric enlarge the training set and deal with the small sample size problem. 4- In contrast to NFLE, which assigns the same weights to all training samples, WFLE assigns different weights to the training samples to rectify their position. The assigned

weights can correct the negative effects of abnormal training samples and so, degrade the classification error. 5- In contrast to NFLE, which uses an initial feature reduction, i.e., PCA, which may loss some directions of data with high discrimination ability, WFLE uses the regularization method to deal with the singularity problem.

The use of FL based methods such as NFLE and the proposed WFLE is not suggested when there is large enough training samples because of two reasons: 1- the computational time exponentially increases as the number of training samples increases , 2- when there is many real training samples, the use of virtual training samples is nonsense.

4 Experiments

In this section, we evaluate the performance of the proposed WFLE method in phishing website detection compared to three different types of supervised feature extraction methods and also with original features of data. The first group of feature extraction methods consists of methods that work based on class discrimination criterion. These methods such as LDA estimate the scatter matrices and try to increase the class separability by maximizing the between-class scatters and minimizing the within-class scatters, and so, they are useful for classification purposes. Because these methods estimate the first and second order statistics of data, they need a large training set to provide the accurate estimation of scatter matrices and so, they have no good efficiency in small sample size situation. But, they are very efficient methods for classification of data when there is a large training set. The second group of feature extraction methods such as LPP consists of methods that preserve the local structure of data. These methods can be implemented either supervised or unsupervised, where the supervised version of them is more appropriate than unsupervised one for classification purposes. Because these methods do not need to estimate the statistics of data, they have low sensitivity to the number of training samples. On the other hand, because they work based on preserving the local structure of data, they are more appropriate for signal representation and sample construction, than classification. However, when there is a limited number of training samples, the use of these methods can be also useful for classification purposes. The third group of feature extraction methods consists of methods that try to enlarge the training set by producing the virtual training samples, and so, they deal with small sample size problem. For an instance, the NFLE method produces the virtual training samples using the FL metric and uses them for estimation of between-class and within-class scatter matrices.

The efficiency of WFLE for phishing website detection is evaluated compared to LDA, supervised LPP, NFLE, and also with the original features. The features extracted by WFLE, LDA, LPP, NFLE, and also the original features are given to a NN classifier with Euclidean distance for classification of phishing website data, i.e., separation of phishing (fake) websites from the legitimate (real) websites.

4.1 Dataset and Evaluation Measures

The unavailability of reliable training datasets for phishing websites is one of the challenges faced by researchers because there is no agreement on definition of discriminative features that characterize the phishing websites. Despite the existence of many articles on predicting a phishing website, using data mining methods, it is yet difficult to collect a dataset that covers all possible features. The used phishing website dataset in this work, which consists of the sound and effective features in predicting phishing websites, is collected mainly from PhishTank archive, MillerSmiles archive, and Google searching operators [35]. This dataset consists of 11055 instances (6157 phishing websites and 4898 legitimate websites) each one containing 30 attributes. The description of attributes of the phishing website dataset is represented in Table 1.

To do experiments, the training samples are chosen randomly from the entire phishing dataset and the remaining samples are considered for testing. Each experiment is done 10 times and the average classification results are represented. The extensive experiments with various training sample size is done to evaluate the performance of the WFLE method in different sizes of training set. The performance of WFLE and other methods are compared with each other using 10, 15, 20, 30, 40, and 50 training samples per class.

Several evaluation measures are used for assessment of any binary classification method such as phishing detection. These measures consist of confusion matrix, phishing classification accuracy (Acc_{phi}), that is also corresponding to true positive rate (TPR) and phishing recall (R), legitimate classification accuracy (Acc_{leg}), false positive rate (FPR), phishing precision (P), F-measure (F), and overall classification accuracy (Acc). Different evaluation measures can be calculated from the confusion matrix where the confusion matrix shows the actual and predicted classification of each class. In this context, positive and negative refer to websites considered as phishing and legitimate, respectively. The number of phishing websites that are correctly detected is denoted by true positive (TP), the number of legitimate websites that

Table 1. The description of attributes of the phishing website dataset.

No.	attribute	No.	attribute
1	having_IP_Address	16	SFH
2	URL_Length	17	Submitting_to_email
3	Shortining_Service	18	Abnormal_URL
4	having_At_Symbol	19	Redirect
5	double_slash_redirecting	20	on_mouseover
6	Prefix_Suffix	21	RightClick
7	having_Sub_Domain	22	popUpWidnow
8	SSLfinal_State	23	Iframe
9	Domain_registeration_length	24	age_of_domain
10	Favicon	25	DNSRecord
11	port	26	web_traffic
12	HTTPS_token	27	Page_Rank
13	Request_URL	28	Google_Index
14	URL_of_Anchor	29	Links_pointing_to_page
15	Links_in_tags	30	Statistical_report

are falsely classified as phishing is denoted by false positive (FP), the number of phishing websites that are falsely classified as legitimate is denoted by false negative (FN), and the number of legitimate websites that are correctly classified is denoted by true negative (TN). As the discrimination threshold between two classified categories is varied, the trade-off between TPR and FPR is described visually by the receiver operating characteristic (ROC) curve. The area under the curve (AUC) is an important scalar measure which is calculated from the ROC curve. A classifier with high AUC value is efficient. The classifiers with AUC values in the range $[0.5, 1]$ are useful and an ideal classifier has AUC value equal to 1. The formulas of used evaluation measures are represented as follows:

$$
\begin{aligned}
Acc_{phi} &= TPR = R = \frac{TP}{TP+FN} \\
Acc_{leg} &= \frac{TN}{FP+TN} \\
FPR &= \frac{FP}{FP+TN} \\
P &= \frac{TP}{TP+FP} \\
F &= \frac{2.P.R}{P+R} \\
Acc &= \frac{TP+TN}{TP+FP+FN+TN}
\end{aligned}
$$

4.2 Experimental Results

The classification results using 10, 15, 20, 30, 40, and 50 training samples are reported in Table 2 and the ROC curves are shown in Figure 7. To avoid the singularity problem, the number of training samples has to be at least one more than the number of attributes (features) of data, i.e., $n \geq d+1$. To assess the effects of the training set size, different number of training samples are used for doing experiments. In the used dataset, we have $d = 30$. So, $n = 10, 15$ training samples can be considered as a small training set; $n = 20, 30$ can be considered as a training set with moderate size and $n = 40, 50$ are associated with a large training set. The method with the highest detection rate and the same false alarm rate has better performance compared to its competitors. In other words, each ROC curve which is the closest to the upper left corner of the coordinates, or the area under it is the largest is associated with the best detection method. The ROC curve that is near to the diagonal of coordinates is approximately associated with a detection method with a 50%-50% detection probability, which is unreliable and undesirable. So, according to the obtained results in Figure 7, the LDA method practically fails to work by using a small number of training samples such as $n = 10, 15$.

According to Figure 7, with $n = 10, 15$, there are obvious gap between the ROC curves of WFLE and other methods especially LDA. So, the superior performance of WFLE by using small training set is observable. By using a training set with moderate size ($n = 20, 30$), the gap between WFLE and other methods is decreased but it is yet significant. By using a large training set ($n = 40, 50$), the ROC curves of WFLE becomes close to other methods. But, it is yet the closest to the upper left corner of the coordinates.

Both of NFLE and WFLE use the virtual training samples obtained by feature line metric. So, both of them can deal with the small sample size problem. So, the performance of them are close together by using a small training set, $n = 15$. From other hand, phishing website datasets contain unreliable and abnormal samples which cause classification error. When the number of available training samples is very low, $n = 10$, the negative effects of abnormal samples are seen more. WFLE, which can deal with the negative effects of abnormal data points through correcting weights, can be superiorly implemented with $n = 10$, while NFLE cannot achieve desirable detection results. So, there is an obvious gap between NFLE and WFLE in very small training sets. From the obtained results, we can conclude the followings:

(1) The proposed WFLE method, almost in all cases, provides the best detection results. It

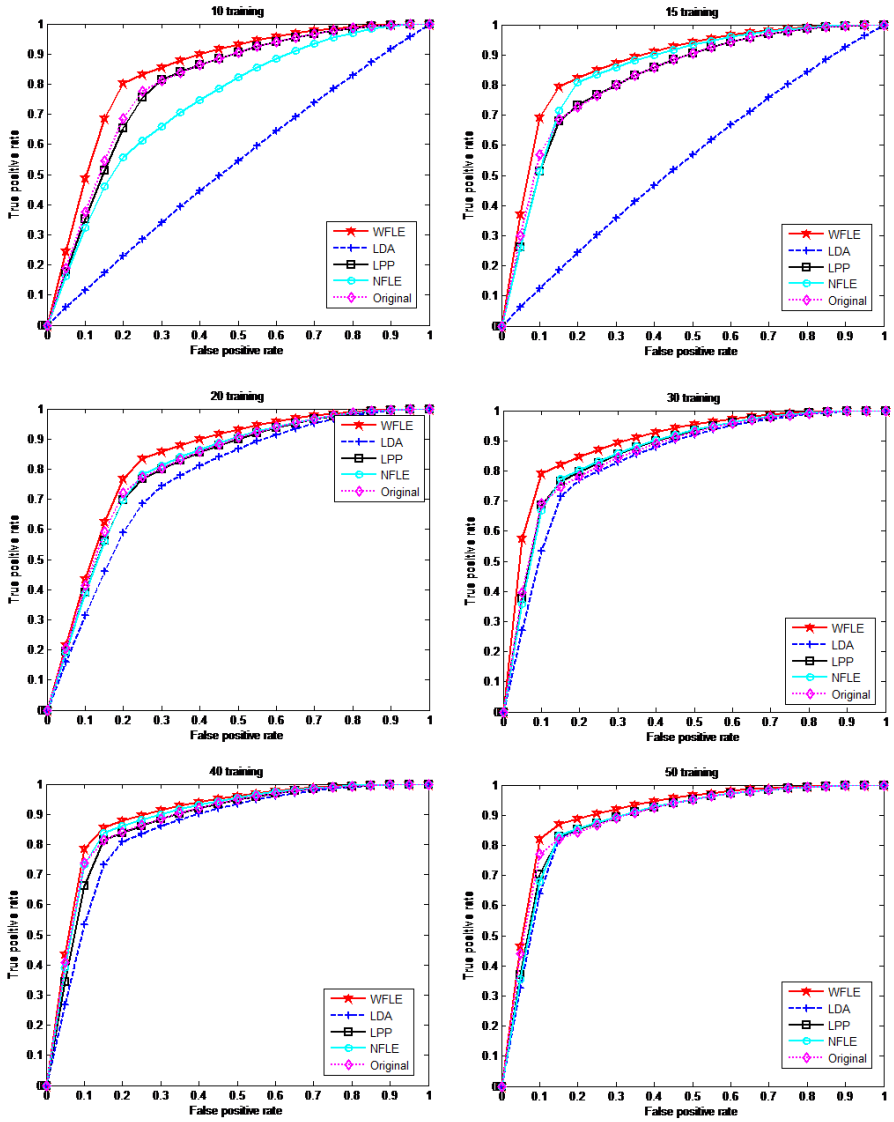

Figure 7. The ROC curves obtained by WFLE, LDA, LPP, NFLE, and original features using 10, 15, 20, 30, 40, and 50 training samples.

is expected, because, it uses the appropriate weights to correct the quality of abnormal samples and so, decreases the classification errors.

(2) Using 10 training samples, the legitimate classification accuracy (Acc_{leg}) and false positive rate (FPR) of LPP is higher than WFLE. This is reasonable because WFLE needs to estimate the first and second order statistics of data to calculate the scatter matrices. Using very small training samples, the accurate estimates of scatter matrices may not be obtained.

(3) In most cases, NFLE is more efficient than LPP and original features. The higher ability of NFLE is because of using virtual training samples.

(4) Original features provide better classification results compared to LPP features in most cases.

(5) LDA is the worst method using very limited training samples. With increasing the number of training samples, the performance of LDA is improved and becomes close to other supervised feature extraction methods.

(6) The efficiency of all methods are improved by increasing the number of training samples, and also the ROC curves are close to the upper left corner of the ROC space.

5 Conclusion

A supervised feature extraction method called WFLE was proposed in this paper to increase the separability between phishing and legitimate websites. In real applications, there is often phishing websites that are very similar to legitimate ones and so, they may wrongly locate in the same category in the feature

Table 2. The classification results obtained by WFLE, LDA, LPP, NFLE, and original features using 10, 15, 20, 30, 40, and 50 training samples.

		$Acc_{phi} = TPR = R$	Acc_{leg}	FPR	P	F	Acc	AUC
10 training samples	**WFLE**	**78.79**	81.23	18.77	**83.93**	**81.28**	**79.88**	**80.01**
	LDA	49.17	55.45	44.55	57.86	53.16	51.97	52.31
	LPP	69.89	**81.40**	**18.60**	82.38	75.62	75.02	75.65
	NFLE	78.51	57.61	42.39	69.73	73.86	69.20	68.06
	Original	72.67	79.67	20.33	81.64	76.89	75.79	76.17
15 training samples	**WFLE**	**86.72**	**78.29**	21.71	83.24	**84.95**	**82.96**	**82.50**
	LDA	40.82	65.92	34.08	59.84	48.53	52.00	53.37
	LPP	83.17	70.67	29.33	77.92	80.46	77.60	76.92
	NFLE	80.18	80.71	**19.29**	**83.79**	81.95	80.42	80.45
	Original	85.95	67.73	32.27	76.82	81.13	77.83	76.84
20 training samples	**WFLE**	**89.92**	**76.30**	**23.70**	**82.51**	**86.06**	**83.85**	**83.11**
	LDA	85.88	66.35	33.65	76.05	80.67	77.18	76.12
	LPP	85.19	72.15	27.85	79.19	82.08	79.38	78.67
	NFLE	85.26	72.66	27.34	79.51	82.28	79.65	78.96
	Original	86.37	70.67	29.33	78.56	82.28	79.38	78.52
30 training samples	**WFLE**	**91.59**	**78.20**	**21.80**	**83.94**	**87.60**	**85.62**	**84.89**
	LDA	82.68	75.09	24.91	80.50	81.58	79.30	78.89
	LPP	87.48	74.83	25.17	81.21	84.23	81.84	81.15
	NFLE	86.65	76.12	23.88	81.87	84.19	81.96	81.39
	Original	88.66	71.80	28.20	79.64	83.91	81.15	80.23
40 training samples	**WFLE**	**87.62**	**84.52**	**15.48**	**87.56**	**87.59**	**86.24**	**86.07**
	LDA	81.08	80.36	19.64	83.70	82.37	80.76	80.72
	LPP	85.05	81.31	18.69	84.99	85.02	83.38	83.18
	NFLE	86.58	83.04	16.96	86.40	86.49	85.00	84.81
	Original	**87.62**	79.84	20.16	84.39	85.98	84.16	83.73
50 training samples	**WFLE**	88.32	**85.64**	**14.36**	**88.44**	**88.38**	**87.12**	**86.98**
	LDA	83.80	83.56	16.44	86.38	85.07	83.69	83.68
	LPP	85.95	82.35	17.65	85.83	85.89	84.35	84.15
	NFLE	85.19	82.96	17.04	86.15	85.66	84.19	84.07
	Original	**88.53**	79.93	20.07	84.58	86.51	84.70	84.23

space. Also, it is possible that two legitimate websites have features that are not very similar together or it is possible that there are two phishing websites that are distant respect together in the feature space. These websites are known as abnormal samples in this work. The proposed WFLE method corrects the undesirable quality of these abnormal samples by using an appropriate weighting manner in the discriminant analysis. Moreover, WFLE produces virtual samples by using FL metric. Thus, it enlarges the training set and deals with the small sample size problem. So, when the number of available labelled websites is limited or the obtained labels are not reliable, WFLE can be an appropriate method for extraction of efficient features to separate phishing from legitimate websites. While soft machine learning methods such as neural network cannot work by using limited number of training samples, WFLE work well in using small training set. For example, the overall classification accuracies obtained by WFLE is 79.88, 82.96,

83.85, 85.62, 86.24 and 87.12 using 10, 15, 20, 30, 40, and 50 training samples per class, respectively.

Acknowledgment

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

References

[1] G. Ramesh, I. Krishnamurthi, K. Sampath Sree Kumar, An efficacious method for detecting phishing webpages through target domain identification, Decision Support Systems 61 (2014) 12-22.

[2] R. Gowtham, I. Krishnamurthi, A comprehensive and efficacious architecture for detecting phishing webpages, computers & security 40 (2014) 23-3 7.

[3] E.-S. M. El-Alfy, A. A. AlHasan, Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm, Future Generation Computer Systems 64 (2016) 98-107.

[4] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, Detecting fake Websites: the contribution of statistical learning theory, MIS Q., 34 (3) (2010), 1-28.

[5] OpenDNS Phishing Quiz. https://www.opendns.com/phishing-quiz/, 2016 (accessed 19.10.16).

[6] APWG Phishing Attack Trends Reports, Retrieved April 21, 2015.

[7] R. M. Mohammad, F. Thabtah, L. McCluskey, Predicting phishing websites based on self-structuring neural network, Neural Comput & Applic., 2013. DOI 10.1007/s00521-013-1490-z.

[8] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based Associative Classification data mining, Expert Systems with Applications, 41 (2014) 5948-5959.

[9] Y. Li, L. Yang, J. Ding, A minimum enclosing ball-based support vector machine approach for detection of phishing websites, Optik, 127 (2016) 345-351.

[10] X. Chen, I. Bose, A. Chung Man Leung, C. Guo, Assessing the severity of phishing attacks: A hybrid data mining approach, Decision Support Systems 50 (2011) 662-672.

[11] M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, Intelligent phishing detection system for e-banking using fuzzy data mining, Expert Systems with Applications 37 (2010) 7913-7921.

[12] V. Ramanathan, H. Wechsler, Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation, computers & security 34 (2013) 123 -139.

[13] N. Abdelhamid, Multi-label rules for phishing classification, Applied Computing and Informatics, Applied Computing and Informatics 11(2015) 29-46.

[14] W. Hadi, F. Aburub, S. Alhawari, A new fast associative classification algorithm for detecting phishing websites, Applied Soft Computing 48 (2016) 729-734.

[15] G. A. Montazer, S. ArabYarmohammadi, Detection of Phishing Attacks in Iranian E-banking Using a Fuzzy-Rough Hybrid System, Applied Soft Computing, 35 (2015) 482-492.

[16] P.A. Barraclough, M.A. Hossain, M.A. Tahir, G. Sexton, N. Aslam, Intelligent phishing detection and protection scheme for online transactions, Expert Systems with Applications 40 (2013) 4697-4706.

[17] D. Zhu, G. Premkumar, X. Zhang, C.-.H. Chu, Data mining for network intrusion detection: a comparison of alternative methods, Decision Sciences 32 (4) (2001) 635-660.

[18] M. Imani, H. Ghassemian, Attribute Profile Based Feature Space Discriminant Analysis for Spectral-Spatial Classification of Hyperspectral Images, Computers and Electrical Engineering, 2016, In Press.

[19] J. Lu, G. Wang, W. Deng and K. Jia, Reconstruction-Based Metric Learning for Unconstrained Face Verification, *IEEE Transactions on Information Forensics and Security*, 10 (1) (2015) 79-89.

[20] N. Martinel, C. Micheloni and G. L. Foresti, Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning, *IEEE Transactions on Image Processing*, 24 (12) (2015) 5645-5658.

[21] H. Wang, L. Feng, J. Zhang and Y. Liu, Semantic Discriminative Metric Learning for Image Similarity Measurement, *IEEE Transactions on Multimedia*, 18 (8) (2016) 1579-1589.

[22] Q. Zhang, L. Zhang, Y. Yang, Y. Tian and L. Weng, Local Patch Discriminative Metric Learning for Hyperspectral Image Feature Extraction, *IEEE Geoscience and Remote Sensing Letters*, 11 (3) (2014) 612-616.

[23] J. Lu, G. Wang and P. Moulin, Localized Multifeature Metric Learning for Image-Set-Based Face Recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, 26 (3) (2016) 529-540.

[24] Y. Wang *et al.*, Learning a Discriminative Distance Metric With Label Consistency for Scene Classification, *IEEE Transactions on Geoscience and Remote Sensing*, 55 (8) (2017) 4427-4440.

[25] K. Fukunaga, Introduction to Statistical Pattern Recognition, San Diego: Academic Press Inc,

1990.
[26] J.-G. Wang, E. Sung,W.-Y. Yau, Incremental two-dimensional linear discriminant analysis with applications to face recognition, Journal of Network and Computer Applications, 33 (2010) 314-322.
[27] X. F. He, P. Niyogi, Locality preserving projections, in Proc. Adv. Neural Inf. Process. Syst. 16 (2004) 153-160.
[28] Y.-L. Chang, J.-N. Liu, C.-C. Han, Y.-N. Chen, Hyperspectral Image Classification Using Nearest Feature Line Embedding Approach, IEEE Trans. Geoscience and remote sensing, 52 (1) (2014) 278-287.
[29] S. Z. Li, J. Lu, Face recognition using the nearest feature line method, IEEE Trans. Neural Netw., 10 (2) (1999) 439-433.
[30] Y. W. Pang, Y. Yuan, X. Li, Generalized nearest feature line for subspace learning, Electron. Lett., 43 (20) (2007) 1079-1080.
[31] J. Lu, Y. P. Tan, Uncorrelated discriminant nearest feature line analysis for face recognition, IEEE Signal Process. Lett., 17 (2) (2010) 185-188.
[32] W.-H. Yang, D.-Q. Dai, Two-Dimensional Maximum Margin Feature Extraction for Face Recognition, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 39 (4) (2009) 1002-1012.
[33] Y.-L. Chang, A simulated annealing feature extraction approach for hyperspectral images, Future Generation Computer Systems 27 (2011) 419-426.
[34] B. C. Kuo and D. A. Landgrebe, Nonparametric weighted feature extraction for classification, IEEE Trans. Geosci. Remote Sens, 42 (5) (2004) 1096-1105.
[35] R. M. Mohammad, L. McCluskey, F. Thabtah, UCI Machine Learning Repository: Phishing Websites Data Set. http://archive.ics.uci.edu/ml/datasets/Phishing+Websites#, 2015 (accessed 2.10.16).

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. Her research interests include pattern recognition, signal and image processing, and remote sensing.

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.