An Efficient Discriminant Analysis Algorithm for Document Classification

Ziqiang Wang and Xia Sun
Henan University of Technology,Zhengzhou,China
Email: wzqagent@126.com

***Abstract*—Document categorization has become one of the most important research areas of pattern recognition and data mining due to the exponential growth of documents in the Internet and the emergent need to organize them. The document space is always of very high dimensionality and learning in such a high dimensional space is often impossible due to the curse of dimensionality. To cope with performance and accuracy problems with high dimensionality, a novel dimensionality reduction algorithm called IKDA is proposed in this paper. The proposed IKDA algorithm combines kernel-based learning techniques and direct iterative optimization procedure to deal with the nonlinearity of the document distribution. The proposed algorithm also effectively solves the so-called "small sample size" problem in document classification task. Extensive experimental results on two real world data sets demonstrate the effectiveness and efficiency of the proposed algorithm.**

***Index Terms*—document classification, kernel discriminant analysis, dimensionality reduction, data mining**

I. INTRODUCTION

With the rapid advances of computer technology and the advent of the World Wide Web, there has been an explosive increase in the amount of document on the Internet. Hence, it is of great importance to develop methods for the automatic processing of large collections of Web documents. One of the main tasks in this processing is that of assigning the documents of a corpus to a set of previously fixed categories, what is known as document classification. Within the last few decades, Document classification (DC) has found a wide range of applications, such as information retrieval, personalized recommendation system, and business intelligence solutions. As a result, numerous DC algorithms have been proposed, and surveys in this area can be found in [1,2].The typical data classification algorithms are directly performed in the data space. However, the document space is always of very high dimensionality, ranging from several hundreds to thousands. Learning in such a high dimensionality in many cases is almost infeasible. Therefore, it is often essential to conduct dimensionality reduction to acquire an efficient and discriminative representation before formally conducting classification. Dimensionality reduction could effectively avoid the "curse of dimensionality", improve performance and computational efficiency of document classification, suppress noise, and alleviate storage requirement. Once the high-dimensional data is mapped into lower-dimensional space, conventional classification algorithms can then be applied.

The most well-known dimensionality reduction methods may be principal component analysis (PCA)[3] and linear discriminant analysis (LDA)[4,5]. Both of them are eigenvector methods aim at modeling linear variability in the multidimensional space. PCA also known as Karhunen–Loéve transformation, aims to find a set of mutually orthogonal bases that capture the global information of the data points in terms of variance. PCA performs dimensionality reduction by projecting the original d-dimensional data onto the r ($<< d$)-dimensional linear subspace spanned by the leading eigenvectors of the data's covariance matrix. By contrast with the unsupervised method of PCA, LDA is a supervised learning approach. LDA seeks a subspace projected onto which the data points of different classes are far away while the data points of the same class are close to each other. LDA aims to find the optimal discriminant vectors by maximizing the ratio of the between-class distance to the within-class distance, thus achieving the maximum class discrimination. It is generally believed that algorithms based on LDA are superior to those based on PCA. The interested reader may refer to [4-6] for detailed analysis about the relationship between PCA and LDA.LDA has been applied successfully in many applications including information retrieval[7], face recognition[8], and microarray data analysis[9].

However, one major drawback of LDA is that it suffers from the small sample size (SSS) or undersampled problem[5]. The small sample size problem arises whenever the number of samples is smaller than the dimensionality of samples. The small sample size problem occurs frequently in practice. For example, in handling document data in information retrieval, it is often the case that the number of terms in the document collection is larger than the total number of documents

This work is supported by the National Natural Science Foundation of China under Grant No.70701013, the Natural Science Foundation of Henan Province under Grant No.0611030100 and 072300430220, and the Natural Science Foundation of Henan University of Technology under Grant No. 08XJC013 and 09XJC016.

Corresponding author: Ziqiang Wang.

doi:10.4304/jsw.6.7.1265-1272

and, therefore, the within-class scatter matrix S_w is singular. To overcome this limitation, many extensions have been proposed to deal with such high-dimensional, undersampled problem, including two-stage PCA+LDA[10], Regularized LDA[11], Penalized LDA[12], Pseudo-inverse LDA[13], Direct LDA[14] , Null space LDA[15,16], Orthogonal LDA[17], Uncorrelated LDA[18], LDA/QR[19] and LDA/GSVD[20] were proposed in the past to deal with the singularity problems. They have been applied successfully in various applications. More details on these methods, as well as their relationship, can be found in [4,17,21].

In addition, although LDA is an efficient linear dimensionality reduction method, it is still a linear technique in nature. So it often fails to find the underlying nonlinear structure of document data sets. Motivated by the kernel trick[22] successfully used in pattern recognition, the classification efficiency induced by LDA may be further improved when the data in the original space are highly nonlinearly distributed. Kernel based nonlinear discriminant analysis algorithms have recently attracted a great deal of attention, these methods are usually called kernel discriminant analysis (KDA) [23]. Their main idea is to transform the input data into a higher dimensional space by a nonlinear mapping function and then apply LDA techniques in that space. These methods are formulated in terms of dot products of the mapped samples, and kernel functions are used to compute these dot products. Therefore, the nonlinear mapping function and the mapped samples are not used explicitly, which makes the methods computationally feasible. KDA performs much better than LDA. Just like LDA, KDA also lead to the small sample size (SSS) problem because the number of the sample is much smaller than the dimension of the representative features of documents. Since SSS problems are common, it is necessary to develop more effective KDA algorithms to deal with them.

Motivated by the kernel trick successfully used in support vector machine (SVM)[22,24], we propose an iterative kernel discriminant analysis (IKDA) method to overcome both the matrix singularity problem and the nonlinear problem for document classification. The IKDA method combines the strengths of both direct iterative optimization procedure and kernel-based learning techniques to improve the performance of LDA. Besides, the proposed IKDA algorithm can effectively solve the so-called small sample size (SSS) problem. We will give detailed derivation of the formulations of IKDA and also make comparison to other conventional kernel-based subspace learning algorithms on a real-world document collection. Experimental results demonstrate the effectiveness and efficiency of our proposed algorithm.

The rest of this paper is organized as follows. The conventional methods for linear and nonlinear discriminant analysis are briefly reviewed in Section II. IKDA algorithm is described in Section III. Experimental results are reported in Section IV. Conclusions are summarized in Section V.

II. Brief review of LDA and KDA

A. Linear Discriminant Analysis(LDA)

LDA is one of the most popular linear dimensionality reduction algorithms. Let $X = \{x_i\}(i = 1,2,\ldots,n)$ be *d*-dimensional sample sets. Each data sample belongs to exactly one of *c* object classes $\{L_1, L_2, \ldots, L_c\}$. n_i denotes the number of samples in class L_i .Thus, $\sum_{i=1}^{c} n_i = n$. LDA seeks a linear transformation matrix $W = (w_1, w_2, \ldots, w_m) \in R^{d\times r}$ mapping the original *d*-dimensional sample space into an *r*-dimensional feature space, where $r << d$ .Then the transformed new feature vectors $y_i \in R^r$ are defined as follows:

$$y_i = W^T x_i , \qquad i = 1,2,\ldots,n \tag{1}$$

LDA seeks directions on which the data samples of different classes are far from each other while requiring data samples of the same class to be close to each other thus achieving the maximum class discrimination. The objective function of LDA is defined as follows:

$$J(W) = \arg\max_W \frac{W^T S_b W}{W^T S_w W} \tag{2}$$

where the between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as follows:

$$S_b = \sum_{i=1}^{c} n_i (u_i - u)(u_i - u)^T \tag{3}$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_j - u_i)(x_j - u_i)^T \tag{4}$$

where x^T represents the transpose of x, u_i the mean of samples in the *i*th class sample set and u the mean of all samples.

$$u_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j , \quad u = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} x_j \tag{5}$$

The optimization problem in (2) is equivalent to finding generalized eigenvectors that correspond to the largest eigenvalues in

$$S_b W = \lambda S_w W \tag{6}$$

The solution can be obtained by solving an eigenvalue problem on the matrix $S_w^{-1} S_b$ if S_w is nonsingular. When the SSS problem takes place, S_w will be typically singular and LDA cannot be applied directly. To overcome this limitation, several extensions, including two-stage PCA+LDA, null space LDA, direct LDA, LDA/QR, LDA/GSVD were proposed in the past to deal

with such singularity problem. A recent overview of LDA on undersampled problems can be found in [17,21].

B. Kernel Discriminant Analysis (KDA)

Although the linearization mapping function of LDA is computationally efficient for classification, its performance may degrade in cases with nonlinearly distributed data. To handle nonlinearly distributed data, LDA is generalized to its kernel version, named as KDA[23]. Its main idea is to transform the input data into a higher dimensional feature space by a nonlinear mapping function and then apply the linear discriminant analysis techniques in the feature space. KDA is capable of handling high-dimensional data and extracting most discriminant features for classification automatically.

To extend the LDA to the nonlinear case, consider a nonlinear feature mapping φ, the input data space $X \subset R^d$ can be mapped into a higher dimensional feature space F.

$$\varphi : R^d \to F, x \mapsto \varphi(x) \tag{7}$$

Without knowing the feature mapping φ explicitly, we can compute dot-products in the feature space F through the following kernel functions:

$$k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{8}$$

As in (3) and (4), the between-class scatter matrix and the within-class S_b scatter matrix S_w in the feature space F are expressed below

$$S_b^{\varphi} = \sum_{i=1}^{c} n_i \left(u_i^{\varphi} - u^{\varphi}\right)\left(u_i^{\varphi} - u^{\varphi}\right)^T \tag{9}$$

$$S_w^{\varphi} = \sum_{i=1}^{c}\sum_{j=1}^{n_i} \left(\varphi(x_j) - u_i^{\varphi}\right)\left(\varphi(x_j) - u_i^{\varphi}\right)^T \tag{10}$$

where $u_i^{\varphi} = (1/n_i)\sum_{j=1}^{n_i} \varphi(x_j)$ and $u^{\varphi} = (1/n)\sum_{i=1}^{c}\sum_{j=1}^{n_i} \varphi(x_j)$ are the mean of the ith class sample set and the mean of all samples in the feature space F, respectively.

According to the theory of reproducing kernels, $W \in F$ must lie in the span of all the training samples in the feature space F .Hence, there exist coefficients $\alpha_i (i = 1,2,\ldots,n)$ such that

$$W = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \tag{11}$$

Using the scatter matrices, the optimal criterion of KDA in the feature space F can be rewritten as

$$J^{\varphi}(W) = \arg\max_{W} \frac{W^T S_b^{\varphi} W}{W^T S_w^{\varphi} W} \tag{12}$$

Substituting (11) into the numerator and denominator of (12), we derive the following equations:

$$W^T S_b^{\varphi} W = \alpha^T K_b \alpha \tag{13}$$

$$W^T S_w^{\varphi} W = \alpha^T K_w \alpha \tag{14}$$

where

$$K_b = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} (m_i - m_j)(m_i - m_j)^T \tag{15}$$

$$K_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i} (\eta_j - m_i)(\eta_j - m_i)^T \tag{16}$$

$$m_i = \left(\frac{1}{n_i}\sum_{j=1}^{n_i} k(x_1, x_j), \frac{1}{n_i}\sum_{j=1}^{n_i} k(x_2, x_j), \ldots, \frac{1}{n_i}\sum_{j=1}^{n_i} k(x_n, x_j)\right)^T$$

$$m_j = \left(\frac{1}{n_j}\sum_{p=1}^{n_j} k(x_1, x_p), \frac{1}{n_j}\sum_{p=1}^{n_j} k(x_2, x_p), \ldots, \frac{1}{n_j}\sum_{p=1}^{n_j} k(x_n, x_p)\right)^T$$

$$\eta_j = \left(k(x_1, x_j), k(x_2, x_j), \ldots, k(x_n, x_j)\right)^T$$

As a result, the solution to (12) can be converted into the following optimization problem:

$$J(\alpha) = \arg\max_{\alpha} \frac{\alpha^T K_b \alpha}{\alpha^T K_w \alpha} \tag{17}$$

Similar with the solution of traditional discriminant analysis, the optimal solution of (17) can also be obtained using the generalized eigenvalue decomposition(GED) method.

$$K_b \alpha_i = \lambda_i K_w \alpha_i \tag{18}$$

where $\lambda_0 \geq \lambda_1 \geq \cdots \lambda_{d-1}$ are the d largest eigenvalues, λ_i is the i-th largest eigenvalue corresponding to eigenvector α_i , and α_i constitutes the i-th column vector of the matrix α .

For a given test data point x , we can compute projections onto the eigenvector W in the feature space F according to

$$\begin{aligned}(W \cdot \varphi(x)) &= \sum_{i=1}^{n} \alpha_i (\varphi(x) \cdot \varphi(x_i)) \\ &= \sum_{i=1}^{n} \alpha_i k(x, x_i)\end{aligned} \tag{19}$$

II. ITERATIVE KERNEL DISCRIMINANT ANALYSIS ALGORITHM

A. Discriminant Feature Extraction

Although KDA performs much better than LDA in many classification applications due to its ability in extracting nonlinear features that exhibit high class separability, it also leads to the small sample size (SSS) problem since the dimensionality of the mapped feature space is usually larger than the size of the training set. To overcome this limitation, the most popular method is to use a penalized term $K_w + \mu I$ instead of K_w in (18), where I is the identity matrix, while it is difficult to determine parameter μ . Lu et al. proposed the kernel

direct LDA (KDDA) method [25],which combined the idea of KDA and direct LDA. KDDA employs the simultaneous diagonalization for finding projection vectors in the range of S_b .However, the range of S_b does not necessarily include the optimal projection vectors for discrimination. Yang developed a two-phase KFD framework[26], which firstly use KPCA to reduce the dimension and then perform standard LDA in the KPCA-transformed space, a limitation of this approach is that the KPCA stage may lose some useful information for discrimination. Recently, Park presented a kernel nonlinear discriminant analysis method using the generalized singular value decomposition (GSVD)[27] to address the singularity problem. However, a disadvantage of this method is the high computational cost of GSVD, especially for large and high-dimensional data sets.

In this paper, inspired from directly solving the trace ratio problem in [28], we propose an efficient iterative kernel discriminant analysis(IKDA) algorithm for directly solving the kernel discriminant analysis problem. Instead of extracting feature vectors from an eigenvalue problem of $K_w^{-1}K_b$ once and for all, the feature vectors will be obtained iteratively. Since no matrix inverse needs to be computed, this algorithm completely avoids the SSS problem. The detailed steps for implementing the IKDA algorithm are summarized as follows.

Step1: Compute kernel matrix K in terms of (8).

Step2: Compute the scatter matrixes K_b and K_w from (15) and (16), respectively.

Step3: Initialize α_0 with a random vector, and normalize it.

Step4: For $t = 1,2,\ldots,t_{\max}$ Repeat

Step4.1: Compute the trace ratio value λ_t from the projection vector α_{t-1}:

$$\lambda_t = \frac{tr\left(\alpha_{t-1}^T K_b \alpha_{t-1}\right)}{tr\left(\alpha_{t-1}^T K_w \alpha_{t-1}\right)} \tag{20}$$

Step4.2: Construct the trace difference problem using the following equation:

$$\alpha_t = \arg\max_{\alpha^T\alpha=1} tr\left(\alpha^T\left(K_b - \lambda_t K_w\right)\alpha\right) \tag{21}$$

Step4.3: Compute the trace difference problem with the eigenvalue decomposition:

$$\left(K_b - \lambda_t K_w\right)\alpha_t(i) = \tau_t(i)\alpha_t(i) \tag{22}$$

where $\tau_t(i)$ is the *i*-th largest eigenvalue of $\left(K_b - \lambda_t K_w\right)$ with the corresponding eigenvector $\alpha_t(i)$, and $\tau_t(0) \geq \tau_t(1) \geq \cdots \tau_t(d-1)$ are the *d* largest eigenvalues.

Step4.4.: If $\left\|\alpha_t - \alpha_{t-1}\right\| < \varepsilon$, then break.

Step4.5: $t = t+1$.

Step5: Output $\alpha = \alpha_t$.

Note that, due to our proposed IKDA algorithm for direct solving KDA is a simplied iterative trace ratio algorithm, it has proofed that the original iterative algorithm converge to a global optimum in [28].As a results, our proposed iterative algorithm IKDA will converge to a global optimum.

B. Classification Method

After the discriminant features are extracted by the IKDA algorithm, a feature matrix is obtained for each document. A remaining key element of document classification is to design a robust classifier. SVM has a very good performance for pattern classification problems by minimizing the Vapnik-Chervonenkis dimensions and achieving a minimal structural risk[22]. Then, the SVM classifier is used for document classification.

Given a set of training document data belonging to two separate classes, $(x_1, y_1),\ldots,(x_n, y_n)$,where x_i denotes the lower-dimensional document feature obtained from the above IKDA algorithm, and $y_i \in \{-1,+1\}$ is the class label, SVM aims to find a hyperplane

$$wx + b = 0 \tag{23}$$

to separate the data, where w is the normal vector to the hyperplane and b is the corresponding bias term of the hyperplane. SVM finds the parameters *w* and *b* for the optimal hyperplane to maximize the geometric margin $2/\|w\|$ subject to

$$y_i\left(w^T x_i + b\right) \geq +1 \tag{24}$$

The above optimization problem can be posed as a constrained quadratic programming (QP) problem, and the solution can be obtained using the Wolfe dual problem with a Lagrangian-multiplier α_i:

$$Q(\alpha) = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \left(x_i \cdot x_j\right) \tag{25}$$

subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

When x_i has a non-zero α_i Lagrange multiplier value, this x_i is called support vector. Only vectors corresponding to nonzero α_i contribute to decision function, and are called support vectors. Thus, the SVM classification function can be derived as:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \left(x_i \cdot x\right) + b\right) \tag{26}$$

where m is the number of support vectors, and b is determined according to the below equation:

$$b = -\tfrac{1}{2} w \cdot \left[x_r + x_s\right] \tag{27}$$

$$w = \sum_{i=1}^{n} \alpha_i x_i y_i \tag{28}$$

where x_r and x_s are any support vector satisfied : $\alpha_r \geq 0, \alpha_s \geq 0, x_r = +1, x_s = -1$.

Although the above linear hyperplane is a natural choice as a boundary to separate different classes, it has limitations for nonlinearly document data. One way to handle nonlinear data can be provided by using kernel trick[22]. The intuition of the kernel trick is to map non-separable data from the original feature space to a higher dimensional Hilbert space

$$\varphi : X \to F \tag{29}$$

in which the data may be linearly separable.

The map φ, rather than being given in an explicit form, is presented implicitly by specifying a kernel function as the inner product between each pair of points in the feature space.

$$x_i \cdot x_j \to \varphi(x_i) \cdot \varphi(x_j) = K(x_i, x_j) \tag{30}$$

where $K(,)$ is a kernel function satisfying Mercer's condition.

Thus, the optimization objective in (25) can be rewritten as follows by using kernel trick:

$$Q(\alpha) = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{31}$$

Finally, the decision function of SVM classifier is given by

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \right) \tag{32}$$

where m is the number of support vectors, each x_i denotes a support vector and α_i is the corresponding Lagrange multiplier.

In this experiment, we adopt the normalized Gaussian kernel as kernel function due to its better performance in many pattern classification applications:

$$K(x_i, x_j) = \frac{k(x_i, x_j)}{\sqrt{k(x_i, x_i)} \cdot \sqrt{k(x_j, x_j)}} \tag{33}$$

where $k(x_i, x_j)$ is Gaussian kernel function $k(x_i, x) = e^{-\rho \|x_i - x\|^2}$, the parameter ρ is set to $\rho = 2^{(n-10)/2.5}\sigma$, where σ is the standard deviation of the data set. In addition, the LIBSVM[29] software was used in our experiment to solve the SVM optimization problem.

Note that, document classification is practically a task of multiclass classification while SVM was designed for the binary classification. While one-against-one(OAO) and one-against-ALL(OAA) schemes are two popular ways to realize the SVM-based multiclass classification task. In this study, we employ the decision-directed acyclic graph (DDAG)[30] learning architecture proposed by Platt et al. to cope with the multiclass classification for his better performance.

In short, the document classification process has three steps. First, we calculate the document subspace from the training set of document data; then the new document to be classified is projected into lower-dimensional feature subspace by using our proposed IKDA algorithm; finally, the new document is classified by the SVM classifier.

III. EXPERIMENTAL RESULTS

In this section, several experiments are carried out to show the efficiency and effectiveness of our proposed IKDA algorithm for document classification. Two standard document collections were used in our experiment: Reuters-21578[31] and 20 Newsgroups[32]. The proposed algorithm is compared with the commonly used kernel-based learning algorithms: KPCA[33], KNDA[34], and KDDA[25]. The Gaussian kernel $K(x, y) = e^{-\rho \|x-y\|^2}$ is used to compute the elements of the matrix $K : k_{ij} = k(x_i, x_j)$, where parameter ρ is selected with leave one out cross validation. Two parameters which control the termination condition of the IKDA algorithm need to be set beforehand: The threshold value ε is simply set to 10^{-4} as in [28]. The maximal iteration number $t_{\max}$ in the IKDA algorithm is experimentally set to 10. All of our experiments have been performed on an Intel P4 3.20GHz PC with 1GB memory.

The Reuters-21578 data set contains 21578 documents in 135 categories from the Reuters newswire[31].Each document belongs to one or more categories. In this experiment, we discarded those documents with multiple category labels, and selected 10 most populated categories (top 10), which are the 10 categories having highest number of documents.

The 20 Newsgroups corpus contains almost 20,000 documents taken from the Usenet newsgroups [32], these documents are evenly distributed on 20 categories.

We simply removed the stop words and no further preprocessing was done. Each document is represented as a term-frequency vector and normalized one. The Euclidean metric is used as the distance measure between document vectors. In addition, in order to remove the uninformative word features, feature selection is conducted using the Information Gain criterion. In particular, top 500 most informative features are selected for each category in each of the two text collections described above. A random subset with *k*(= 5%, 10%, 20%, 30%, 40%, 50%) samples per category are selected for training and the rest are used for testing.

To evaluate the effectiveness of the document classification algorithm, the average classification accuracies and the running time (second) of computing each dimensionality reduction algorithm on two data sets are listed on the Table (I-IV). To reduce the variability, for given the percent of training samples in each class, we average the resulting accuracies of 10 random splits and report the mean value. From the experimental results, we can make the following observations.

1) The proposed IKDA algorithm consistently outperforms KPCA, KNDA, and KDDA in terms of classification accuracy. It implies that the numerical computation problem does affect the performance of kernel-based discriminant analysis algorithm. IKDA successfully avoids the numerical computation problem since it does not calculate any inverse matrix for delivering discriminant features.

2) The proposed IKDA algorithm is more efficient than KPCA, KNDA, and KDDA in terms of running time. The reason it that the discriminant feature vectors of IKDA algorithm are obtained through iterative trace ratio calculation, instead of direct computing matrix inverse which is computationally expensive.

3) These kernel-based LDA algorithms(such as: KNDA, KDDA, and IKDA) give a relatively better performance compared with kernel-based PCA(KPCA) algorithm. One possible explanation is as follows: As the same as PCA, KPCA captures the overall variance of all features which is optimal for pattern representation, not necessarily for classification. In addition, the higher computational complexity of KPCA is due to the used significantly larger feature number.

4) Although KNDA, KDDA, and IKDA algorithm belong to kernel-based discriminant analysis algorithm, the proposed IKDA algorithm performs much better than KNDA and KDDA. The reasons are listed the following: (a) The discarded null space by the KNDA algorithm may contain the most significant discriminant information.(b) KDDA computes the discriminant vectors in the orthogonal complement of the null space of the between-class scatter matrix. However, neither in the null space of the between-class scatter matrix nor in its orthogonal complement includes the most discriminant vectors in the null space of the within-class scatter matrix. Hence, the KDDA algorithm may fail to find these important discriminant vectors.

In addition, in order to test the significance of the improvement obtained by our proposed IKDA algorithm, we did a *t*-test at the significance level (i.e., 0.05) on the classification accuracy among different algorithms given percent of training samples in each class. The test results are shown in Table VI and Table VII. For the Reuters-21578 and 20 Newsgroups data sets, all of these tests demonstrate that the performance of our proposed IKDA algorithm outperforms KPCA significantly ($p<0.05$). Although the classification accuracy rate of IKDA is still better than that of KNDA and KDDA, the performance difference between them is not statistically significant.

IV. CONCLUSIONS

A new document classification algorithm called IKDA has been introduced in this paper. The proposed IKDA algorithm combines kernel-based learning techniques and direct iterative optimization procedure to provide an efficient and effective approach for improving the performance of LDA. In addition, this algorithm completely avoids the SSS problem since no matrix inverse needs to be computed. Extensive experiments on Reuters-21578 and 20 Newsgroups data sets demonstrate that IKDA is superior to related algorithms in terms of effectiveness and efficiency. An open problem in IKDA is the selection of kernel function and its parameters, which is also an unsolved problem in kernel-based learning algorithm. We are currently studying this problem in theory and practice.

TABLE I.
CLASSIFICATION ACCURACY ON 20 NEWSGROUPS

Size	IKDA	KPCA	KDDA	KNDA
1%	74.36%	65.28%	72.43%	73.54%
5%	79.58%	67.32%	77.58%	77.42%
10%	82.61%	70.48%	78.21%	80.28%
20%	86.27%	74.87%	82.37%	83.31%
30%	89.53%	78.75%	83.28%	84.46%
40%	90.43%	81.36%	84.75%	85.96%
50%	91.23%	84.72%	85.15%	87.23%

TABLE II.
CLASSIFICATION ACCURACY ON REUTERS-21578

Size	IKDA	KPCA	KDDA	KNDA
1%	80.23%	67.14%	79.32%	80.21%
5%	85.98%	70.25%	83.28%	83.98%
10%	89.37%	75.63%	85.42%	85.42%
20%	91.69%	81.28%	85.51%	86.35%
30%	95.45%	83.31%	90.38%	91.41%
40%	96.68%	86.27%	93.65%	93.64%
50%	98.53%	88.38%	94.89%	95.28%

TABLE III.
RUNNING TIME ON REUTERS-21578(S)

Size	IKDA	KPCA	KDDA	KNDA
1%	12.53	15.78	13.67	13.74
5%	14.78	18.85	15.52	15.82
10%	17.72	23.23	18.17	17.94
20%	21.65	28.19	24.12	23.71
30%	26.81	33.43	28.65	27.42
40%	30.74	37.75	34.96	32.73
50%	35.12	42.25	28.34	36.76

TABLE VI.
RUNNING TIME ON 20 NEWSGROUPS (S)

Size	IKDA	KPCA	KDDA	KNDA
1%	12.63	14.85	13.35	13.82
5%	15.58	18.23	15.78	16.37
10%	18.35	22.45	21.46	20.29
20%	21.76	25.72	23.53	22.58
30%	25.82	29.35	28.24	27.61
40%	30.46	33.21	32.35	33.73
50%	35.87	38.48	36.48	37.42

TABLE IV.
P VALUES OF T-TEST ON REUTERS-21578 BETWEEN DIFFERENT ALGORITHMS

Algorithm	KPCA	KDDA	KNDA
IKDA	0.00885	0.28622	0.35604

TABLE V.
P VALUES OF T-TEST ON 20 NEWSGROUPS BETWEEN DIFFERENT ALGORITHMS

Algorithm	KPCA	KDDA	KNDA
IKDA	0.0165	0.17033	0.3234

REFERENCES

[1] F.Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*,vol.34,pp.1-47, March 2002.

[2] C.J.van Rijsbergen, *Information Retrieval*, 2nd edition. London:Butterworths,1979.

[3] I.T.Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag,1986.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition.New York:Academic,1990.

[5] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Second Edition. Hoboken:Wiley-Interscience,2000.

[6] X.Wang and X.Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.26,pp.1222-1228,Septmber 2004.

[7] K.Torkkola,"Discriminative features for text document classification," *Pattern Analysis & Applications*, vol.6, pp. 301-308,February 2004.

[8] Q.Liu,H.Lu,and S.Ma, "Improving kernel fisher discriminant analysis for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.14,pp.42-49,January 2004.

[9] J.Ye, T.Li, T.Xiong, and R.Janardan,"Using uncorrelated discriminant analysis for tissue classification with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*,vol.1,pp.181-190,October-December 2004.

[10] P.N.Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19,pp.711-720,July 1997.

[11] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol.84, pp.165-175, March 1989.

[12] T. Hastie and R. Tibshirani, "Penalized discriminant analysis," *Annals of Statistics*, vol.23,pp.73-102,January 1995.

[13] M.Skurichina and R.P.W. Duin, "Stabilizing classifiers for very small sample size," *Proceedings of the 13th International Conference on Pattern Recognition*,pp.891-896,August 1996.

[14] H.Yu and J. Yang,"A direct LDA algorithm for high-dimensional data-with application to face recognition," *Pattern Recognition*, vol.34,pp.2067–2070,October 2001.

[15] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*,vol.33,pp.1713-1726,October 2000.

[16] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," *Proceedings of the 16th International Conference on Pattern Recognition*,pp.29-32,December 2002.

[17] J.Ye, and T.Xiong, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *Journal of Machine Learning Research*, vol.7, pp.1183-1204,July 2006.

[18] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis," *Proceedings of the twenty-first international conference on Machine learning*,pp.113-120,July 2004.

[19] J.Ye and Q.Li, "LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation," *Pattern Recognition*,vol.37,pp.851-854,April 2004.

[20] P.Howland and H.Park,"Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol.26,pp.995-1006,August 2004.

[21] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis undersampled problems," *Journal of Machine Learning Research*, vol.6, pp.483-502,April 2005.

[22] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer,1995.

[23] G.Baudat and F.Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol.12, pp.2385-2404,October 2000.

[24] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkof, "An introduce to kernel based learning algorithm," *IEEE Transactions on Neural Networks*, vol.12, pp.181-201.August 2002.

[25] J.Lu, K.N.Plataniotis, and A.N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis

algorithms," *IEEE Transactions on Neural Networks*, vol.14, pp.117-126, January 2003.
[26] J.Yang, A.F. Frangi,Y.Yang,D.Zhang, and Z.Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol.27, pp.230-244,February 2005.
[27] C.H.Park and H.Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*,vol.27,pp.87-102,January 2005.
[28] H.Wang,S.Yan,D.Xu,X.Tang, and T.Huang, "Trace ratio vs. ratio trace for dimensionality reduction," *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*,pp.1-8,June 2007.
[29] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm,2001.
[30] J.C.Platt, N.Cristianini, and J. Shawe-Taylor,"Large margin DAGs for multiclass classification," *Advances in Neural Information Processing Systems* 12, pp.547-553, November 1999.
[31] Reuters-21578. http://www.daviddlewis.com /resources/ testcollections/reuters21578/, 2004.
[32] 20 News Group. http://kdd.ics.uci.edu/databases/ 20newsgroups/20newsgroups.htm, 2004.
[33] B. Scholkopf, A. Smola, and K.Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *MPI fur biologische kybernetik*, Tubingen, Germany, Technology Report 44, 1996.
[34] W. Liu, Y. H.Wang, S. Z. Li, and T.N.Tan, "Null space-based kernel Fisher discriminant analysis for face recognition," *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp.369-374, May 2004.

Ziqiang Wang was born Zhoukou, Henan, China in 1973. He received the PhD degree from Xi'an Jiaotong University and the Master of Science degree from Xi'an Petroleum University, Xi'an, China, both in computer science, in 1999 and 2005, respectively.

He is currently an associate professor in the Henan University of Technology, Zhengzhou, China. His research interests include data mining and pattern recognition..

Xia Sun was born Xi'an, Shanxi, China in 1978. She received the Master of Science degree in computer application from Huazhong University of Science and Technology, Wuhan, China, in 2008.

She is currently a lecturer in the Henan University of Technology, Zhengzhou, China. Her research interests include data mining and pattern recognition