Boosting Kernel Discriminant Analysis and Its Application to Tissue Classification of Gene Expression Data

Guang Dai & Dit-Yan Yeung
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{daiguang,dyyeung}@cse.ust.hk

Abstract

Kernel discriminant analysis (KDA) is one of the most effective nonlinear techniques for dimensionality reduction and feature extraction. It can be applied to a wide range of applications involving high-dimensional data, including images, gene expressions, and text data. This paper develops a new algorithm to further improve the overall performance of KDA by effectively integrating the boosting and KDA techniques. The proposed method, called *boosting kernel discriminant analysis* (BKDA), possesses several appealing properties. First, like all kernel methods, it handles nonlinearity in a disciplined manner that is also computationally attractive; second, by introducing pairwise class discriminant information into the discriminant criterion and simultaneously employing boosting to robustly adjust the information, it further improves the classification accuracy; third, by calculating the significant discriminant information in the null space of the within-class scatter operator, it also effectively deals with the small sample size problem which is widely encountered in real-world applications for KDA; fourth, by taking advantage of the boosting and KDA techniques, it constitutes a strong ensemble-based KDA framework. Experimental results on gene expression data demonstrate the promising performance of the proposed methodology.

1 Introduction

Principal component analysis (PCA) and linear discriminant analysis (LDA) are two classical feature extraction and dimensionality reduction techniques broadly used in many tasks involving high-dimensional data. It is generally believed that for pattern classification problems such as face recognition and tissue classification of gene expression data, LDA-based algorithms usually outperform PCA-based ones. The reason is that the former optimizes the low-dimensional representation of the objects for classification by maximizing the ratio of the between-class scatter to the within-class scatter, while the latter simply optimizes object reconstruction without taking class information into consideration. Many LDA-based algorithms have been proposed. Among them, as an attractive approach, some new LDA algorithms [Lu *et al.*, 2003a; Masip and Vitrià, 2006] which effectively integrate the boosting and LDA techniques have been developed recently to further enhance the classification performance.

On the other hand, a major problem of linear subspace methods such as PCA and LDA is that they fail to extract nonlinear features representing higher-order statistics. In order to overcome this limitation, kernel dimensionality reduction techniques such as kernel principal component analysis (KPCA) [Schölkopf *et al.*, 1999] and kernel discriminant analysis (KDA) [Baudat and Anouar, 2000; Lu *et al.*, 2003b; Dai and Qian, 2004; Xiong *et al.*, 2005; Dai and Yeung, 2005; Zheng *et al.*, 2005] have been proposed recently to extend linear dimensionality reduction techniques to nonlinear versions by using the kernel trick, demonstrating better performance in many applications than that obtained using their linear counterparts. The basic idea is to first map each input data point $x \in \mathbb{R}^n$ into the feature space $\mathcal{F}$ via a nonlinear mapping ϕ and then apply the corresponding linear dimensionality reduction algorithm in $\mathcal{F}$. Moreover, similar to their linear counterparts, KDA-based methods are generally better than KPCA-based methods for classification tasks. However, KDA-based algorithms usually suffer from the small sample size problem, because the number of training examples available is usually smaller than the dimensionality of the feature space $\mathcal{F}$ especially for high-dimensional data. In order to overcome this problem, many approaches have been developed based on different criteria. Recently, more effective solutions [Dai and Qian, 2004; Zheng *et al.*, 2005] have been proposed to calculate the optimal discriminant vectors in the null space of the within-class scatter operator where the significant discriminant information exists. On the other hand, similar to conventional LDA-based techniques [Lotlikar and Kothari, 2000; Tang *et al.*, 2005; Dai and Yeung, 2005], the performance of KDA degrades further due to the following deficiencies which are referred to as non-balanced problems in this paper:

1. For multi-class pattern classification problems, the optimal criterion based on the conventional between-class scatter is not directly related to classification accuracy. In particular, the corresponding dimensionality reduction procedure tends to overemphasize the inter-class distances of well-separated outlier classes in the input space at the expense of classes that are close to each other, leading to significant overlap between them.
2. The expression of the average within-class scatter has an implicit assumption that all classes have the same weight

for the covariances. In fact, if the class with dominant covariance is simultaneously an outlier class in the input space, the within-class scatter will fail to estimate the correct value for improved classification, due to minimizing the spread of the outlier classes while neglecting the minimization of other covariances.

In this paper, we further improve the overall performance of KDA by proposing a novel KDA algorithm called *boosting KDA* (BKDA). The proposed approach effectively integrates the boosting technique with the most recently developed KDA algorithms based on the pairwise class discriminant information. The BKDA approach employs the boosting technique to robustly calculate the pairwise class discriminant information that is integrated into the scatter operators in order to solve the non-balanced problems of KDA. Although boosting has been applied to LDA, so far it has not been studied much for KDA. It is worthwhile to mention here several appealing properties of the proposed BKDA method:

1. It handles nonlinearity in a disciplined manner that is also computationally attractive like all kernel methods.
2. By introducing the pairwise class discriminant information into the discriminant criterion and simultaneously employing boosting to robustly adjust the information, it effectively overcomes the non-balanced problems in KDA and further increases the classification accuracy.
3. It effectively boosts the significant discriminant information contained in the null space of the within-class scatter operator and simultaneously deals with the small sample size problem. However, the algorithm and analysis presented here can also be applied easily to boost other discriminant information such as that included in the orthogonal complement of the null space of the within-class scatter operator.
4. It constitutes a strong ensemble-based kernel KDA framework, taking advantage of both the boosting and KDA techniques.

To demonstrate the effectiveness of the BKDA method, we apply the proposed method to effectively extract discriminant features for the tissue classification of gene expression data. Experimental results confirm that our BKDA method is superior to several existing dimensionality reduction methods in terms of classification accuracy.

2 Nonlinear Feature Extraction via Kernel Discriminant Analysis

As a nonlinear extension of LDA, KDA essentially performs LDA in the feature space $\mathcal{F}$. Note that $\mathcal{F}$ may be infinite-dimensional and it should be regarded as a Hilbert space. As such, the scatter matrices in the input space for LDA correspond to operators in $\mathcal{F}$ for KDA. For any operator $\mathbf{A}$ in $\mathcal{F}$, we let $\mathbf{A}(0)$ denote the null space of $\mathbf{A}$, i.e., $\mathbf{A}(0) = \{\mathbf{x} \,|\, \mathbf{A}\mathbf{x} = \mathbf{0}\}$, and $\overline{\mathbf{A}(0)}$ the orthogonal complement of $\mathbf{A}(0)$. Thus, $\mathbf{A}(0) \oplus \overline{\mathbf{A}(0)} = \mathcal{F}$.

Suppose $\mathcal{X}$ denotes a training set of N examples belonging to C classes, with each example being a vector in $\mathbb{R}^n$. Let $\mathcal{X}_i \subset \mathcal{X}$ be the ith class containing N_i examples, with $\mathbf{x}_i^{i_o}$ denoting the i_oth example in it. In addition, each example $\mathbf{x}_i^{i_o} \in \mathcal{X}$ has the corresponding label $i \in \{1, \ldots, C\}$. By the implicit nonlinear mapping $\phi : \mathbb{R}^n \to \mathcal{F}$, the N images in $\mathcal{F}$ can be represented by the set $\{\phi(\mathbf{x}_i^{i_o})\}$. The conventional between-class scatter operator $\mathbf{S}_b^\phi$, within-class scatter operator $\mathbf{S}_w^\phi$, and population scatter operator $\mathbf{S}_t^\phi$ can be expressed as: $\mathbf{S}_b^\phi = \frac{1}{N}\sum_{i=1}^{C} N_i(\mathbf{m}_i^\phi - \mathbf{m}^\phi)(\mathbf{m}_i^\phi - \mathbf{m}^\phi)^T$, $\mathbf{S}_w^\phi = \frac{1}{N}\sum_{i=1}^{C}\sum_{i_o=1}^{N_i}(\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}_i^\phi)(\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}_i^\phi)^T$, $\mathbf{S}_t^\phi = \mathbf{S}_b^\phi + \mathbf{S}_w^\phi = \frac{1}{N}\sum_{i=1}^{C}\sum_{i_o=1}^{N_i}(\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}^\phi)(\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}^\phi)^T$, where $\mathbf{m}_i^\phi = \frac{1}{N_i}\sum_{i_o=1}^{N_i}\phi(\mathbf{x}_i^{i_o})$ and $\mathbf{m}^\phi = \frac{1}{N}\sum_{i=1}^{C}\sum_{i_o=1}^{N_i}\phi(\mathbf{x}_i^{i_o})$. We maximize the Fisher criterion below to obtain the optimal projection directions $\mathbf{w}$ in $\mathcal{F}$:

$$J^\phi(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_b^\phi\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w^\phi\mathbf{w}}. \tag{1}$$

However, in many practical applications, one of the major problems with KDA is the so-called small sample size problem with respect to (1), because of the degeneracy of the within-class scatter operator in $\mathcal{F}$. In general, this problem is solved by applying techniques such as pseudoinverse, kernel PCA, and QP decomposition. Recently, some more effective methods [Dai and Qian, 2004; Zheng *et al.*, 2005] have been developed to explore the significant discriminant information in the null space of the within-class scatter operator. In general, a simple and efficient approach of finding this significant discriminant information is to calculate the optimal discriminant vectors in the intersection subspace $\overline{\mathbf{S}_t^\phi(0)}\bigcap\mathbf{S}_w^\phi(0)$ with respect to the following modified criterion:

$$J_f^\phi(\mathbf{w}) = \mathbf{w}^T\mathbf{S}_b^\phi\mathbf{w} \ (\|\mathbf{w}\| = 1). \tag{2}$$

Furthermore, in many situations, nonlinear feature extraction based on the above criterion (2) indeed further enhances the overall performance in terms of classification accuracy and numerical stability. In addition, it should be pointed out that, according to [Chen *et al.*, 2000], for conventional LDA, when the dimensionality n of the training examples is less than the total number of examples N, some useful discriminant information in the null space of the within-class scatter operator will be lost. This is especially the case when n is less than $N - C$ because all this discriminant information will be fully discarded. In KDA described above, this situation can be avoided by choosing the kernel function appropriately, with which the dimensionality of $\mathcal{F}$ nonlinearly mapped from the input space can become larger than the number of training examples.

3 Boosting Kernel Discriminant Analysis Learner

3.1 Boosting Kernel Discriminant Analysis

Similar to LDA, KDA also suffers from the non-balanced problems described in Section 1. Here, we will present the BKDA algorithm that combines the strengths of the boosting and KDA techniques and effectively solves the non-balanced problems of KDA at the same time.

Boosting is a general machine learning meta-algorithm for improving the accuracy of any given learning algorithm. One of the most effective boosting algorithms, referred to as AdaBoost, can be used in conjunction with many learning algorithms to improve their performance. Boosting algorithms

are adaptive in the sense that the classifiers built are tweaked in favor of those instances misclassified by previous classifiers [Freund and Schapire, 1997]. More specifically, the underlying idea of AdaBoost is based on the sample distribution, which, in essence, is a measure of how hard to classify an example. Moreover, it should be notable that for multi-class problems, a version of AdaBoost called AdaBoost.M2 outperforms AdaBoost.M1 for many real-world applications. Thus we prefer using AdaBoost.M2 in this paper. In order to effectively overcome the non-balanced problems of KDA and simultaneously form a strong connection between KDA and AdaBoost.M2, following [Freund and Schapire, 1997; Lu *et al.*, 2003a], the pairwise class discriminant distribution is introduced on the basis of the mislabel distribution from AdaBoost.M2. Further, at the tth iteration in AdaBoost.M2, any pairwise class discriminant distribution $d_{i,j}^t$ between classes $\mathcal{X}_i$ and $\mathcal{X}_j$ can be calculated as follows:

$$d_{i,j}^t = \begin{cases} \frac{1}{2}\left(\sum_{i_o=1}^{N_i} \Gamma^t(\mathbf{x}_i^{i_o}, j) + \sum_{j_o=1}^{N_j} \Gamma^t(\mathbf{x}_j^{j_o}, i)\right), & \text{if } i \neq j; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Here, the mislabel distribution $\Gamma^t(\star, *)$ measures the extent of the difficulty of discriminating the example $\star$ from the improper label $*$ on the basis of the previous boosting results. Obviously, a larger value of $d_{i,j}^t$ intuitively indicates worse separability between two classes $\mathcal{X}_i$ and $\mathcal{X}_j$, further embodying also that both are closer together in $\mathcal{F}$.

To address the first non-balanced problem in KDA, we replace the ordinary between-class scatter operator $\mathbf{S}_b^\phi$ by a weighted between-class scatter operator $\mathbf{S}_B^\phi$:

$$\mathbf{S}_B^\phi = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \frac{N_i N_j}{N^2} w(d_{i,j}^t)(\mathbf{m}_i^\phi - \mathbf{m}_j^\phi)(\mathbf{m}_i^\phi - \mathbf{m}_j^\phi)^T, \quad (4)$$

where the weighting function $w(\star)$ is generally chosen to be a monotonically increasing function of $\star$. For simplicity, we let $w(\star) = \star$ in this paper. Obviously, based on the definitions of the pairwise class discriminant distribution and the weighting function $w(\star)$, it can be seen that classes that are not well separated in $\mathcal{F}$ and thus can potentially impair the classification performance should be more heavily weighted in $\mathcal{F}$.

In addition, a weighting scheme should also be employed to alleviate the second non-balanced problem, i.e., the negative effect of outlier classes when estimating the within-class scatter operator. Further, at the tth iteration, we replace the ordinary within-class scatter operator $\mathbf{S}_w^\phi$ by the weighted within-class scatter operator $\mathbf{S}_W^\phi$ as follows:

$$\mathbf{S}_W^\phi = \frac{1}{N} \sum_{i=1}^{C} \sum_{i_o=1}^{N_i} r_i^t q_{ii_o}^t (\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}_i^\phi)(\phi(\mathbf{x}_i^{i_o}) - \mathbf{m}_i^\phi)^T, \quad (5)$$

where $r_i^t = \sum_{j \neq i} w(d_{i,j}^t)$ is the relevance-based weight for class $\mathcal{X}_i$ and $q_{ii_o}^t = w(\sum_{j \neq i} \Gamma^t(\mathbf{x}_i^{i_o}, j))$ is that for $\mathbf{x}_i^{i_o}$. Let us highlight some characteristics of $\mathbf{S}_W^\phi$ here:

- By incorporating r_i^t in (5), it ensures that the estimated $\mathbf{S}_W^\phi$ is only influenced slightly if class $\mathcal{X}_i$ is an outlier class. This is reasonable since if one class is well separated from the other classes in $\mathcal{F}$, then whether the within-class covariance operator of this class in the new space is compact or not will not have much effect on classification [Tang *et al.*, 2005].
- Generally, a larger value of $q_{ii_o}^t$ indicates a greater difficulty of classifying example $\mathbf{x}_i^{i_o}$ w.r.t. the previous boosting results. As such, $q_{ii_o}^t$ emphasizes the difficult examples in the within-class covariance of class $\mathcal{X}_i$.

Based on the AdaBoost.M2 algorithm and the definitions of $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$, we propose the BKDA algorithm as detailed in Fig.1.[1] It should be pointed out that the KDA involved in the BKDA algorithm is to calculate the optimal discriminant vectors in the null space of the new $\mathbf{S}_W^\phi$ in order to take advantage of the significant discriminant information in $\mathbf{S}_W^\phi(0)$, and the detailed calculation procedure can be found in subsection 3.2 below. In addition, discriminant analysis is quite a strong feature extraction technique for classification. As a result, the boosting process cannot go forward due to the very small pseudo-loss ϵ_t. In general, some sampling procedures are employed to artificially weaken the corresponding discriminant technique, and, in BKDA, we choose some examples in each class based on $q_{ii_o}^t$ to focus on the hardest examples in each class. For the features extracted by discriminant analysis, a simple nearest neighbor classifier is generally employed for classification. In order to be consistent with the AdaBoost algorithm, in BKDA, the hypothesis $h_t(\star, *)$ between example $\star$ and class $*$ can be built easily based on the normalized nearest neighbor classifier, while it gives results identical to the classical nearest neighbor classifier.

3.2 How to Calculate the Optimal Discriminant Vectors in the Feature Space $\mathcal{F}$

In what follows, we will describe how to efficiently calculate the significant discriminant information in the null space of the variant $\mathbf{S}_W^\phi$ in (5). Furthermore, on the basis of the variants $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$, we also represent the corresponding population scatter operator as $\mathbf{S}_T^\phi = \mathbf{S}_B^\phi + \mathbf{S}_W^\phi$. As described in Section 2, the optimal discriminant vectors can be obtained in $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$ with respect to the following criterion:

$$J_F^\phi(\mathbf{w}) = \mathbf{w}^T \mathbf{S}_B^\phi \mathbf{w} \ (\|\mathbf{w}\| = 1). \quad (6)$$

Furthermore, based on the above criterion, we have to calculate $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$, which requires calculating $\overline{\mathbf{S}_T^\phi(0)}$ or $\mathbf{S}_W^\phi(0)$ first. However, the computation of $\overline{\mathbf{S}_T^\phi(0)}$ or $\mathbf{S}_W^\phi(0)$ is quite intractable to some extent, due to the following reasons: 1) unlike $\mathbf{S}_t^\phi$, it is intractable to directly compute $\overline{\mathbf{S}_T^\phi(0)}$ by the eigenanalysis of $\mathbf{S}_T^\phi$, since $\mathbf{S}_T^\phi$ cannot be explicitly expressed as $\mathbf{S}_T^\phi = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ is an operator with the explicit formulation; 2) the direct computation of $\mathbf{S}_W^\phi(0)$ is very infeasible, since $\mathbf{S}_W^\phi(0)$ is in general very large in $\mathcal{F}$; 3) according to [Cevikalp *et al.*, 2005], $\mathbf{S}_W^\phi(0)$ can be indirectly calculated by $\mathcal{F} - \overline{\mathbf{S}_W^\phi(0)}$ on the basis of first calculating $\overline{\mathbf{S}_W^\phi(0)}$, while, in $\mathcal{F}$, high computational complexity is involved at the same time. In order to efficiently solve this problem, we provide the following theorem.

[1] For simplicity, KDA in subsection 3.2 is described on $\mathcal{X}$ but it is essentially very similar to that based on $\mathcal{S}_t$.

Input: A set of training examples $\mathcal{X} = \{\mathbf{x}_i^{i_o} \,|\, \mathbf{x}_i^{i_o} \in \mathbb{R}^n, i = 1,\ldots,C, i_o = 1,\ldots,N_i\}$; a set of all mislabels $\mathcal{M} = \{[(i,i_o),j] \,|\, i,j \in \{1,\ldots,C\}, i_o \in \{1,\ldots,N_i\}, i \neq j\}$; the initial mislabel distribution on $\mathcal{M}$: $\Gamma^t(\mathbf{x}_i^{i_o},j) = \frac{1}{|\mathcal{M}|} = \frac{1}{N(C-1)}$; a small constant ε.
for $t = 1,\ldots,T_{max}$ **do**

- Calculate the terms $d_{i,j}^t$ by (3), and r_i^t and $q_{ii_o}^t$ in (5).
- Select s hardest examples per class based on $q_{ii_o}^t$ to form a training subset $\mathcal{S}_t \subset \mathcal{X}$.
- Apply KDA in subsection 3.2 on $\mathcal{S}_t$, and constitute the KDA-based feature extraction technique, denoted by KDA-t; apply KDA-t on $\mathcal{X}$ to obtain $\{\mathbf{y}_i^{i_o,t} \in \mathbb{R}^r\}$; build the hypothesis $h_t(\star,*) \in [0,1]$ on the subset $\mathcal{Y}^t$ of $\{\mathbf{y}_i^{i_o,t} \in \mathbb{R}^r\}$, corresponding to $\mathcal{S}_t$.
- Calculate the pseudo-loss based on h_t: $\epsilon_t = \sum_{[(i,i_o),j]\in\mathcal{M}} \frac{\Gamma^t(\mathbf{x}_i^{i_o},j)(1+h_t(\mathbf{y}_i^{i_o,t},j)-h_t(\mathbf{y}_i^{i_o,t},i))}{2}$.
- Set $\beta_t = \epsilon_t/(1-\epsilon_t)$. If $\beta_t \leq \varepsilon$, then $T_{max} = t-1$ and break.
- Update the mislabel distribution Γ^t: $\Gamma^{t+1}(\mathbf{x}_i^{i_o},j) = \Gamma^t(\mathbf{x}_i^{i_o},j)\beta_t^{(1+h_t(\mathbf{y}_i^{i_o,t},i)-h_t(\mathbf{y}_i^{i_o,t},j))/2}$.
- Normalize Γ^{t+1}: $\Gamma^{t+1}(\mathbf{x}_i^{i_o},j) = \Gamma^{t+1}(\mathbf{x}_i^{i_o},j)/(\sum_{[(l,l_o),g]\in\mathcal{M}} \Gamma^{t+1}(\mathbf{x}_l^{l_o},g))$.

end for
Output: The final hypothesis: $h_f(\mathbf{x}) = \sup_{i\in\{1,\ldots,C\}}\{-\sum_{t=1}^{T_{max}} \log(\beta_t) h_t(\mathbf{y}^t,i)\}$, where, for any example $\mathbf{x}$, $\mathbf{y}^t \in \mathbb{R}^r$ is the corresponding nonlinear feature vector extracted by KDA (KDA-t) in subsection 3.2 over $\mathcal{S}_t$.

Figure 1: Summary of BKDA algorithm.

Theorem 1. *The subspace* $\overline{\mathbf{S}_T^\phi(0)} = \{\mathbf{x} \,|\, \langle \mathbf{S}_T^\phi, \mathbf{x}\rangle \neq 0, \mathbf{x} \in \mathcal{F}\}$ *is equivalent to the subspace* $\overline{\mathbf{S}_t^\phi(0)} = \{\mathbf{x} \,|\, \langle \mathbf{S}_t^\phi, \mathbf{x}\rangle \neq 0, \mathbf{x} \in \mathcal{F}\}$, *where* $\mathbf{S}_T^\phi = \mathbf{S}_B^\phi + \mathbf{S}_W^\phi$ *and* $\mathbf{S}_t^\phi$ *is the conventional population scatter operator.*

Based on Theorem 1, we thus propose to use the following steps to calculate the optimal discriminant vectors for (6): 1) Calculate the orthonormal basis of $\overline{\mathbf{S}_t^\phi(0)}$; 2) Calculate the orthonormal basis of $\mathbf{S}_W^\phi(0)$ in $\overline{\mathbf{S}_t^\phi(0)}$ to construct $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$; 3) Calculate the optimal discriminant vectors with respect to the discriminant criterion (6) in $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$.

In what follows, we will present the detailed computation procedure. From the discussions above, we first need to calculate the orthonormal basis of $\overline{\mathbf{S}_t^\phi(0)}$ by applying KPCA on $\mathbf{S}_t^\phi$. Moreover, $\mathbf{S}_t^\phi$ can be rewritten as:

$$\mathbf{S}_t^\phi = \frac{1}{N}\sum_{i=1}^{C} \Phi_i \Phi_i^T = \mathbf{\Xi}\mathbf{\Xi}^T, \tag{7}$$

where $\Phi_i = [\phi(\mathbf{x}_i^1) - \mathbf{m}^\phi, \ldots, \phi(\mathbf{x}_i^{N_i}) - \mathbf{m}^\phi]$ and $\mathbf{\Xi} = \frac{1}{\sqrt{N}}[\Phi_1,\ldots,\Phi_C]$. Since the dimensionality of $\mathbf{S}_t^\phi$ in $\mathcal{F}$ is usually very high or even infinite-dimensional, KPCA can be carried out by the eigenanalysis of $\mathbf{\Xi}^T\mathbf{\Xi}$ instead, with size $N\times N$. From the training set $\{\phi(\mathbf{x}_i^{i_o})\}$, an $N\times N$ matrix $\mathbf{K}$ can be defined as $\mathbf{K} = (\mathbf{K}_{ij})_{i,j=1,\ldots,C}$ where $\mathbf{K}_{ij} = (k_{i_o j_o})_{i_o=1,\ldots,N_i}^{j_o=1,\ldots,N_j}$ and $k_{i_o j_o} = \langle \phi(\mathbf{x}_i^{i_o}), \phi(\mathbf{x}_j^{j_o})\rangle$. By the kernel trick, $\mathbf{\Xi}^T\mathbf{\Xi}$ can be expressed as

$$\mathbf{\Xi}^T\mathbf{\Xi} = \frac{1}{N}\left[\mathbf{K} - \frac{1}{N}(\mathbf{K1} + \mathbf{1K}) + \frac{1}{N^2}\mathbf{1K1}\right], \tag{8}$$

where $\mathbf{1}$ is an $N\times N$ matrix with all terms being one. Let λ_l and $\mathbf{e}_l$ $(l = 1,\ldots,m)$ be the ith positive eigenvalue and the corresponding eigenvector of $\mathbf{\Xi}^T\mathbf{\Xi}$, respectively. Then $\boldsymbol{\theta}_l = \mathbf{\Xi}\mathbf{e}_l\lambda_l^{-1/2}$ $(l = 1,\ldots,m)$ constitute the orthonormal basis of $\overline{\mathbf{S}_t^\phi(0)}$ (or $\overline{\mathbf{S}_T^\phi(0)}$) [Dai and Qian, 2004; Zheng *et al.*, 2005]. Hence, any input $\mathbf{x} \in \mathbb{R}^n$ can be transformed into $\mathbf{r}$ in the low-dimensional space $\mathbb{R}^m$ via KPCA as follows:

$$\mathbf{r} = \mathbf{P}^T\phi(\mathbf{x}) = \sqrt{1/N}\mathbf{E}^T(\mathbf{I} - \mathbf{1}/N)^T\mathbf{\Upsilon}^T, \tag{9}$$

where $\mathbf{P} = [\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m]$, $\mathbf{E} = [\mathbf{e}_1\lambda_1^{-1/2},\ldots,\mathbf{e}_m\lambda_m^{-1/2}]$, $\mathbf{I}$ is the identity matrix, and $\mathbf{\Upsilon} = (k_1^1,\ldots,k_1^{N_1}, k_2^1,\ldots,k_2^{N_2},\ldots,k_C^1,\ldots,k_C^{N_C})$ with $k_i^{i_o} = \langle\phi(\mathbf{x}_i^{i_o}), \phi(\mathbf{x})\rangle$. Then, by (9), the training examples $\{\mathbf{x}_i^{i_o}\}$ in $\mathbb{R}^n$ (or $\{\phi(\mathbf{x}_i^{i_o})\}$ in $\mathcal{F}$) can be mapped to the corresponding points $\{\mathbf{r}_i^{i_o}\}$ in the low-dimensional space $\mathbb{R}^m$. The scatter matrices $\mathbf{S}_B$ and $\mathbf{S}_W$ in the low-dimensional space $\mathbb{R}^m$ corresponding to $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$ can be computed by either of the following two approaches (for simplicity, this paper adopts the first approach):

- By the corresponding definitions of $\mathbf{S}_B^\phi$ and $\mathbf{S}_W^\phi$, $\mathbf{S}_B$ and $\mathbf{S}_W$ can be reconstructed based on the training examples $\{\mathbf{r}_i^{i_o}\}$ in $\mathbb{R}^m$;
- By the kernel trick, the scatter matrices can be directly computed using $\mathbf{S}_B = \mathbf{P}^T\mathbf{S}_B^\phi\mathbf{P}$ and $\mathbf{S}_W = \mathbf{P}^T\mathbf{S}_W^\phi\mathbf{P}$, similar to KPCA.

Furthermore, we can directly calculate the null space of $\mathbf{S}_W^\phi$ in $\overline{\mathbf{S}_T^\phi(0)}$ by the eigenanalysis of $\mathbf{S}_W$, since $\mathbf{S}_W$ is only an $m\times m$ matrix. More specifically, let $\mathbf{V} = [\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_u]$ be the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_W$, and then $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$ can be spanned by $\mathbf{PV}$. As a result, the discriminant criterion (6) can be further transformed into the projection space of $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$ by $J_F(\mathbf{z}) = \mathbf{z}^T\mathbf{V}^T\mathbf{P}^T\mathbf{S}_B^\phi\mathbf{PVz} = \mathbf{z}^T\mathbf{V}^T\mathbf{S}_B\mathbf{Vz}\,(\|\mathbf{z}\| = 1)$. Let $\mathbf{z}_l$ $(l = 1,\ldots,r)$ be the eigenvectors of $\mathbf{V}^T\mathbf{S}_B\mathbf{V}$, sorted in descending order of the corresponding eigenvalues. According to [Dai and Qian, 2004; Zheng *et al.*, 2005], it is clear that $\mathbf{w}_l = \mathbf{PVz}_l$ $(l = 1,\ldots,r)$ constitute the optimal discriminant vectors with respect to the corresponding criterion (6) in $\overline{\mathbf{S}_T^\phi(0)} \bigcap \mathbf{S}_W^\phi(0)$. For an input pattern $\mathbf{x}$, its corresponding nonlinear feature vector extracted by the KDA procedure described above can be computed as $\mathbf{y} = \langle\mathbf{W}, \phi(\mathbf{x})\rangle \in \mathbb{R}^r$, where $\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_r)$. This expression can be rewritten via the kernel trick as follows:

$$\mathbf{y} = \langle\mathbf{W}, \phi(\mathbf{x})\rangle = \sqrt{1/N}(\mathbf{z}_1,\ldots,\mathbf{z}_r)^T\mathbf{V}^T\mathbf{E}^T(\mathbf{I} - \mathbf{1}/N)^T\mathbf{k}_x.$$

Here, $\mathbf{1}$ is an $N\times N$ matrix with all terms being one, $\mathbf{I}$ is the identity matrix, and $\mathbf{k}_x = (k(\mathbf{x},\mathbf{x}_1^1),\ldots,k(\mathbf{x},\mathbf{x}_C^{N_C}))^T$.

4 Experimental Results

Gene expression data are usually high-dimensional involving a large number of genes but a small sample size. Thus, how to effectively extract discriminant features plays a very important role in gene expression data classification [Ye *et al.*, 2004]. To evaluate the performance of the proposed BKDA algorithm, we conduct some gene expression classification experiments to compare BKDA with some other dimensionality reduction methods.

Since the non-balanced problems mentioned above only exist in multi-class classification problems, our experiments are performed on three different data sets involving more than two classes each in order to demonstrate the behavior of BKDA:[2]

1. 11_Tumors: 174 human tumor examples corresponding to 11 different cancer types;
2. A subset **A** of 14_Tumors: contains all 218 human tumor examples corresponding to 14 different cancer types;
3. A subset **B** of 14_Tumors: contains all examples corresponding to various human tumor and normal tissue types, each of which contains at least eight examples.

Each data set is randomly partitioned into disjoint training and test sets. For the training set $\mathcal{X}$, each class $\mathcal{X}_i$ contains L examples. Since the first data set is less challenging than the second and third data sets, we set $L = 5$ for the first data set and $L = 7$ for the other two. Moreover, in our experiments, no preprocessing such as gene selection is applied on the data sets. For each feature extraction method, we use a simple and efficient minimum mean distance rule with Euclidean distance measure to assess the classification accuracy, and simultaneously build the corresponding normalized version for the hypothesis in BKDA based on [Lu *et al.*, 2003a]. Each experiment is repeated 10 times and the average classification rate is reported. For the kernel methods, we use the RBF kernel $k(\mathbf{z}_1, \mathbf{z}_2) = \exp(\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\sigma)$ and polynomial kernel $k(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{z}_1^T \mathbf{z}_2/\sigma + 1)^2$ where $\sigma = 10^{12}$. In addition, it should be noted that BKDA can effectively boost the discrimination ability of the different features extracted by KDA in BKDA. To reduce the computational cost and simultaneously consider the space limitation in this paper, we simply fix the number of features r in BKDA to $C - 3$ for each data set, where C is the number of classes.

For comparison, the first set of experiments implements BKDA and KDA [Zheng *et al.*, 2005] using both the polynomial and RBF kernels above, since KDA [Zheng *et al.*, 2005] can effectively calculate the significant discriminant information in the null space of the within-class scatter operator and the proposed method [Ye *et al.*, 2004] for gene expression data classification is in essence a special case of KDA in high-dimensional spaces. Furthermore, to explicitly show the effectiveness of BKDA, in each comparison, KDA offers two baselines: KDA* denotes the maximum classification rate over the variant features; KDA-r denotes the classification rate based on r fixed features. In addition, in BKDA, the number of chosen examples per class is set to $s = L - 1$, where L is the number of examples for each class in the training set. The experimental results shown in Fig. 2 reveal that: as expected, BKDA is capable of improving the overall performance for both kernel functions on all three data sets due to its advantages as discussed above; specially, BKDA can effectively improve the performance of the corresponding KDA-r on the same number of features. In addition, we find that when s is fixed at very small values, BKDA fails to show its effectiveness when compared with KDA* since the ensemble-based learner is too weak.

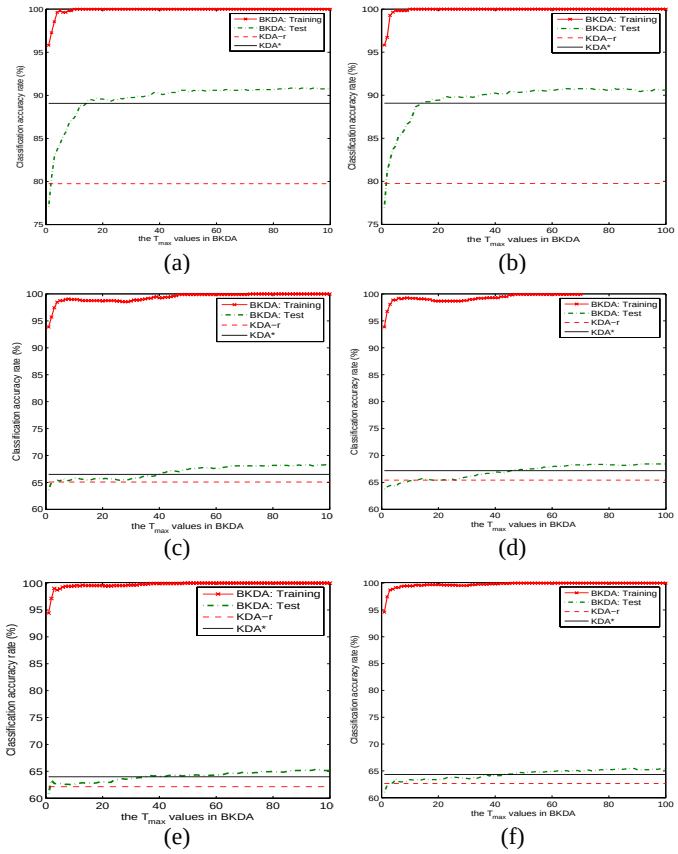

Figure 2: Comparative performance of BKDA and KDA under different T_{max} values for BKDA. BKDA:Training and BKDA:Test denote the results of BKDA on the training set $\mathcal{X}$ and test set, respectively. (a) Polynomial kernel on 11_Tumors; (b) RBF kernel on 11_Tumors; (c) Polynomial kernel on subset **A** of 14_Tumors; (d) RBF kernel on subset **A** of 14_Tumors; (e) Polynomial kernel on subset **B** of 14_Tumors; (f) RBF kernel on subset **B** of 14_Tumors.

The second experiment compares BKDA with several effective linear dimensionality reduction methods, including BDLDA [Lu *et al.*, 2003a], LPP [He and Niyogi, 2004], and NPE [He *et al.*, 2005], and several other effective kernel-based nonlinear dimensionality reduction methods, including KPCA [Schölkopf *et al.*, 1999], GDA [Baudat and Anouar, 2000], KDDA [Lu *et al.*, 2003b], WKDA [Dai and Yeung, 2005], and KDA/QR [Xiong *et al.*, 2005]. Table 1 reports the maximum classification rates of different methods on the three data sets. We can see that BKDA is generally more effective than all other methods compared here.

Notice that there has been extensive research based on classifiers for tissue classification of gene expression data [Statnikov *et al.*, 2005], including k-nearest neighbor classifier

[2] All data sets are available at http://discover1.mc.vanderbilt.edu /discover/public/mc-svm/.

Table 1: Maximum classification rates (%) of different methods on three data sets.

Algorithm	11_Tumors		subset **A**		subset **B**	
	Poly.	RBF	Poly.	RBF	Poly.	RBF
KPCA	78.74	78.82	52.92	52.83	49.63	49.49
GDA	80.84	80.67	63.33	64.08	59.71	60.15
KDDA	84.29	84.12	63.75	64.17	60.07	60.37
WKDA	89.08	89.08	66.50	67.17	63.97	64.34
KDA/QR	84.87	84.12	63.67	64.25	60.06	61.69
Our BKDA	90.84	90.76	68.33	68.42	65.37	65.44
BDLDA	86.55		65.17		63.90	
LPP	83.78		57.58		55.81	
NPE	84.79		61.42		58.01	

(kNN), decision tree, naive Bayes classifier, bagging, boosting, support vector machine (SVM), etc. However, our comparative experiments have not included these methods since we focus on the dimensionality reduction and feature extraction aspect in our study. In addition, over the past decade or so, dimensionality reduction techniques have been playing a very important role in face recognition. We have also applied BKDA to face recognition based on the extended YaleB and PIE databases, showing again that BKDA outperforms all other dimensionality reduction methods compared. The results are not included in this paper due to space limitation.

5 Conclusion

In this paper, a novel KDA algorithm has been presented by incorporating the boosting technique into KDA to give the BKDA algorithm. This new algorithm effectively integrates the strengths of the boosting and KDA techniques to give an ensemble-based KDA framework with strong nonlinear feature extraction capability, and simultaneously overcomes the non-balanced and small sample size problems commonly encountered by KDA-based methods. Extensive empirical comparison of BKDA with many other linear and nonlinear dimensionality reduction methods on gene expression data classification shows that BKDA is a very promising method.

6 Acknowledgments

This research has been supported by Competitive Earmarked Research Grant 621305 from the Research Grants Council of the Hong Kong Special Administrative Region, China.

References

[Baudat and Anouar, 2000] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

[Cevikalp *et al.*, 2005] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):4–13, January 2005.

[Chen *et al.*, 2000] 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*, 33(10):1713–1726, 2000.

[Dai and Qian, 2004] G. Dai and Y.T. Qian. Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2697–2700, 2004.

[Dai and Yeung, 2005] G. Dai and D.Y. Yeung. Nonlinear dimensionality reduction for classification using kernel weighted subspace method. In *Proceedings of the IEEE International Conference on Image Processing*, pages 838–841, September 2005.

[Freund and Schapire, 1997] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[He and Niyogi, 2004] X.F. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, USA, 2004.

[He *et al.*, 2005] X.F. He, D. Cai, S.C. Yan, and H.J. Zhang. Neighborhood preserving embedding. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1208–1213, October 2005.

[Lotlikar and Kothari, 2000] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):623–627, June 2000.

[Lu *et al.*, 2003a] J.W. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Boosting linear discriminant analysis for face recognition. In *Proceedings of the IEEE International Conference on Image Processing*, pages 657–660, 2003.

[Lu *et al.*, 2003b] J.W. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, 14(1):117–126, January 2003.

[Masip and Vitrià, 2006] D. Masip and J. Vitrià. Boosted discriminant projections for nearest neighbor classification. *Pattern Recognition*, 39(2):164–170, 2006.

[Schölkopf *et al.*, 1999] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1999.

[Statnikov *et al.*, 2005] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(1):631–643, 2005.

[Tang *et al.*, 2005] E.K. Tang, P.N. Suganthan, X. Yao, and A.K. Qin. Linear dimensionality reduction using relevance weighted lda. *Pattern Recognition*, 38(4):485–493, 2005.

[Xiong *et al.*, 2005] T. Xiong, J.P. Ye, Q. Li, V. Cherkassky, and R. Janardan. Efficient kernel discriminant analysis via QR decomposition. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, USA, 2005.

[Ye *et al.*, 2004] J.P 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*, 1(4):181–190, 2004.

[Zheng *et al.*, 2005] W. Zheng, L. Zhao, and C. Zou. Foley-Sammon optimal discriminant vectors using kernel approach. *IEEE Transactions on Neural Networks*, 16(1):1–9, January 2005.