Robust and Accurate Cancer Classification with Gene Expression Profiling

Haifeng Li
Dept. of Computer Science
University of California
Riverside, CA 92521
hli@cs.ucr.edu

Keshu Zhang
Human Interaction Research Lab
Motorola, Inc.
Tempe, AZ 85282
keshu.zhang@motorola.com

Tao Jiang
Dept. of Computer Science
University of California
Riverside, CA 92521
jiang@cs.ucr.edu

Abstract

Robust and accurate cancer classification is critical in cancer treatment. Gene expression profiling is expected to enable us to diagnose tumors precisely and systematically. However, the classification task in this context is very challenging because of the curse of dimensionality and the small sample size problem. In this paper, we propose a novel method to solve these two problems. Our method is able to map gene expression data into a very low dimensional space and thus meets the recommended samples to features per class ratio. As a result, it can be used to classify new samples robustly with low and trustable (estimated) error rates. The method is based on linear discriminant analysis (LDA). However, the conventional LDA requires that the within-class scatter matrix $\mathbf{S}_w$ *be nonsingular. Unfortunately,* $\mathbf{S}_w$ *is always singular in the case of cancer classification due to the small sample size problem. To overcome this problem, we develop a generalized linear discriminant analysis (GLDA) that is a general, direct, and complete solution to optimize Fisher's criterion. GLDA is mathematically well-founded and coincides with the conventional LDA when* $\mathbf{S}_w$ *is nonsingular. Different from the conventional LDA, GLDA does not assume the nonsingularity of* $\mathbf{S}_w$*, and thus naturally solves the small sample size problem. To accommodate the high dimensionality of scatter matrices, a fast algorithm of GLDA is also developed. Our extensive experiments on seven public cancer datasets show that the method performs well. Especially on some difficult instances that have very small samples to genes per class ratios, our method achieves much higher accuracies than widely used classification methods such as support vector machines, random forests, etc.*

1 Introduction

Accurate diagnosis of human cancer is essential in cancer treatment. Recently, the advances in microarray technology enable us to simultaneously observe the expression levels of many thousands of genes on the transcription level. In principle, tumor gene expression profiles can serve as molecular fingerprints for cancer classification. Researchers believe that gene expression profiling could be a precise, objective, and systematic method for cancer classification [19, 26, 33]. Many classifiers have been applied to cancer classification, such as nearest neighbor, artificial neural networks, support vector machines, boosting, weighted voting, etc. [5, 9, 10, 11, 17, 19, 26, 31, 33, 38, 45, 47].

Although gene expression profiling provides a great opportunity for accurate and objective cancer diagnosis, the classification task in this context is very challenging because of the very high dimensionality of data since gene expression data usually involve thousands of genes. The high dimensionality is a major practical limitation facing many pattern recognition technologies, especially when the number of samples is small. In practice, it has been observed that a large number of features may degrade the performance of classifiers if the number of training samples is small relative to the number of features [24, 35, 36]. This fact, which is referred to as the "peaking phenomenon", is caused by the "curse of dimensionality" [4]. It is generally accepted that one needs at least 5 – 10 times as many training samples per class as the number of features to obtain well-trained (robust) classifiers [14, 24, 35]. In the case of cancer classification, the number of tumor samples is usually only several dozen due to limitations on sample availability, identification, acquisition, time, and cost. On the other hand, the dimensionality (number of genes) is many thousands. Consequently, dimensionality reduction is essential to cancer classification.

In the past several decades, many dimension reduction techniques have been proposed. Roughly, these methods follow two approaches, *feature selection* and *feature extraction*. Feature selection methods choose a "best" subset of features from a large initial set, taking into account both the cost of selection and the effectiveness of each feature in the classification process. The advantage of feature selection is that the selected features retain their original biological or physical interpretation. So, the retained features help us understand patterns more precisely, and even find the biological/physical process that generates the patterns. On the other hand, feature extraction techniques transform the original feature space into a reduced feature space that has fewer dimensions with little reduction on the effectiveness of classifier. Feature extraction generally provides a better discriminative ability than feature selection. However, the new features generated by (nonlinear) feature extraction may not have a clear biological/physical meaning.

Currently, almost all gene-expression-based cancer classification systems employ some feature selection method for dimension reduction. Ideally, one should select the expression levels of tumor-specific genes for classification. However, only few tumor-specific molecular markers are currently known by molecular oncology. Instead, we have to select marker genes computationally. Of the applied feature selection methods, the single-gene-rank approach is the most common. This approach ignores the interdependence among genes and ranks the relative class separation of each feature independently. The top ranked genes are selected for training the classifier. For example, Golub *et al.* used the signal-to-noise (S2N) metric ratio $P(g) = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}$ to rank genes, where μ_1 and μ_2 are the mean expression levels of gene g in classes 1 and 2, respectively, and σ_1 and σ_2 are the standard deviations of expression levels in classes 1 and 2, respectively [19]. Of course, such univariate feature selection methods are not optimal because a subset of top k ranked genes is not guaranteed to be the best among all subsets of k genes. In particular, many genes are coexpressed due to the complicated genetic networks. Thus, the expression of genes is not independent and it is very hard for univariate feature selection methods to capture the joint discriminant capability of genes.

To improve single-gene-ranking, Bø and Jonassen suggested gene-pair-ranking that simultaneously analyzes pairs of genes to decide the subset of marker genes [6]. Clearly, it is not sufficient to investigate the joint discriminant capability of only a pair of genes. More generally, subsets of $K > 2$ features should be considered. However, it is not practical to jointly select a subset of genes in a brute-force way. For example, the number of ways to select 50 elements from 2000 elements is approximately 10^{100}. Instead, some heuristic has to be applied. In [28, 29], for instance, Li *et al.* proposed the GA/KNN method for gene selection. First, GA/KNN finds many (random) subsets of $K > 2$ genes of expected classification power using a Genetic Algorithm (GA). The "fitness" of each subset of genes is determined by its ability to classify the training set samples according to the k-nearest neighbor (kNN) method. When many such subsets of genes are obtained, the frequencies with which genes are selected are analyzed. The most frequently selected genes are presumed to be the most relevant to sample distinction and are finally used for prediction. Although GA/KNN avoids brute-force search, it is still much slower than univariate feature selection and our proposed method (to be discussed later). The user also has to determine many parameters of the algorithm, such as chromosome length, the number of chromosomes, termination metric, etc.

Due to its popularity and success in many application areas, some researchers have used support vector machine (SVM) to select features. For example, the one-dimensional SVM method ranks genes by the accuracies of single-gene SVM classifiers [40]. This is actually an univariate feature selection method and does not exhibit superiority to other univariate methods [30]. Note that, any other classifiers may be employed instead of SVM in this approach. Another feature selection method based on SVM is recursive feature elimination (RFE) [33]. RFE recursively removes features based on the absolute magnitude of the hyperplane elements of trained SVM. In RFE, the SVM is trained with all genes at first. The expression values of genes whose absolute value of corresponding hyperplane element is in the bottom 10% are removed. Then, the SVM is retrained with the selected genes. This procedure is repeated iteratively to study prediction accuracy as a function of the gene number. It was reported that RFE achieves the similar results as the signal-to-noise ratio method [19] and radius-margin-ratio method [31, 44] on the GCM dataset [33].

It was observed that no matter what feature selection method is employed, at least 50 (and frequently more) features would need be chosen and used for classification in general [39]. This number is quite far from the recommended 5 – 10 times ratio of samples to features per class for the training of a robust classifier, and thus the estimated error rate could be greatly biased [14, 24, 35]. Hence, it is not clear how well the previously proposed methods perform if large datasets are available.

In this paper, we propose a novel linear feature extraction method for dimension reduction in cancer classification based on linear discriminant analysis (LDA). After dimension reduction, a template matching procedure is employed for classification. LDA can map the data into the discriminant space with a very low dimensionality of $c - 1$, where c is the number of classes. For instance, the mapped space is one dimensional for binary classification. So, the mapped data meet the recommended 5 – 10 times ratio of samples to features per class and thus even a small number of samples are sufficient to train a good classifier. Therefore, our method is more robust than others and the estimated error rates are more accurate and trustable. Although the method sounds straightforward, there is a big challenge. Namely, the conventional LDA cannot be applied when the within-scatter matrix $\mathbf{S}_w$ is singular due to the *small sample size* *problem* [16]. The small sample size problem arises when the number of samples is smaller than the dimensionality of samples, which always happens in cancer classification. To overcome the small sample size problem, we propose a generalized linear discriminant analysis (GLDA) by carefully investigating the properties of scatter matrices $\mathbf{S}_b$, $\mathbf{S}_w$ and $\mathbf{S}_t$. GLDA is mathematically well-founded and coincides with the conventional LDA when $\mathbf{S}_w$ is nonsingular. Different from the conventional LDA, GLDA does not assume the nonsingularity of $\mathbf{S}_w$, and thus naturally solves the small sample size problem. To deal with the high dimensionality of scatter matrices, we also develop a fast algorithm for GLDA based on singular value decomposition (SVD). Our experimental results show that the method performs well. Especially on some difficult instances that have very small samples to genes per class ratios, our method achieves much higher accuracies than widely used classification methods such as support vector machines, random forests, etc.

The rest of the paper is organized as follows. Section 2 introduces LDA and our method for cancer classification. Section 3 discusses the small sample size problem and presents our GLDA. In this section, we also develop a fast algorithm for GLDA to accommodate the high dimensionality of gene expression data. In Section 4, we describe some experimental results on several public cancer gene expression datasets in comparison with many other methods. Section 5 concludes the paper with some directions of further research.

2 Linear Discriminant Analysis and Cancer Classification

Before, a feature extraction method, principal component analysis (PCA), had been applied for dimensionality reduction in cancer classification [26]. PCA is a linear mapping that minimizes the mean squared error criterion [25]. PCA computes the d largest eigenvectors of the covariance matrix of D-dimensional samples. The d largest eigenvectors that constitute the mapping matrix are also called as the principal components. These few principal components describe most of the variance of the data. By expanding the data on these orthogonal principal components, we have the minimal reconstruction error. However, the decorrelation and high measures of statistical significance provided by the first few principal components are no guarantees of revealing the class structure that we need for proper classification. As done in many face recognition systems [42], d is usually set to $n-1$ for no information loss when the sample size is much smaller than the dimensionality, where n is the number of samples. However, such a setting cannot meet the recommended 5 – 10 times ratio of samples to features per class. Besides, the fact that the category information associated with samples is neglected also implies that PCA may be significantly suboptimal.

Instead of PCA, we employ linear discriminant analysis (LDA, also called Fisher's Linear Discriminant) [12, 34], which is a supervised feature extraction (and classification) method. In many applications, LDA has proven to be very powerful and performs much better than PCA. Besides, LDA can map the data into the discriminant space of dimensionality $c-1$ so that we can meet the recommended 5 – 10 times ratio of samples to features per class. Thus, we can classify tumors more robustly. Recall that LDA is given by a linear transformation matrix $\mathbf{W} \in \mathcal{R}^{D\times d}$ maximizing the so-called Fisher criterion (a kind of *Rayleigh* coefficient) [12, 34]

$$J(\mathbf{W}) = \text{tr}\left(\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_w\mathbf{W}}\right) \qquad (1)$$

where $\mathbf{S}_b = \sum_{i=1}^{c} p_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ and $\mathbf{S}_w = \sum_{i=1}^{c} p_i E[(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T|\mathcal{C}_i] = \sum_{i=1}^{c} p_i \mathbf{\Sigma}_i$ are the between-class scatter matrix and the within-class scatter matrix, respectively; c is the number of classes; $\mathbf{m}_i$ and p_i are the mean vector and *a priori* probability of class i, respectively; $\mathbf{m} = \sum_{i=1}^{c} p_i\mathbf{m}_i$ is the overall mean vector; $\mathbf{\Sigma}_i$ is the covariance matrix of class i; D and d are the dimensionalities of the data before and after the transformation, respectively; and tr denotes the trace of a square matrix, i.e. the sum of the diagonal elements. Besides, the total/mixture scatter matrix, i.e. the covariance matrix of all samples regardless of their class assignments, is defined as $\mathbf{S}_t = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T] = \mathbf{S}_w + \mathbf{S}_b$ [16]. In LDA, the transformation matrix $\mathbf{W}$ is constituted by the largest eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ to maximize (1) by assuming the nonsingularity of $\mathbf{S}_w$. Since $\mathbf{S}_w^{-1}\mathbf{S}_b$ has at most $d = \min(c-1, D)$ non-zero eigenvalues, $\mathbf{W}$ is usually constituted by the corresponding eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$. For cancer classification, d is at most $c-1$ since the number of classes is always less than the dimensionality. For a new sample $\mathbf{x}$, the predicted class is

$$\mathcal{C}(\mathbf{x}) = \arg\min_k \sum_{i=1}^{d} (\mathbf{w}_i^T(\mathbf{x}-\mathbf{m}_k))^2 \qquad (2)$$

i.e. the class whose mean vector is closest to $\mathbf{x}$ in the discriminant space. Note that for binary classes, one can easily adjust the above decision rule in order to prefer sensitivity or specificity.

As a linear feature extraction method, LDA could also be used to identify important marker genes for further investigation. After training, the elements of each column vector in the mapping matrix of LDA can be thought of as the weights of genes, which determine the importance of genes in classification. By ranking the absolute values of the elements of column vectors in the mapping matrix, we can select marker genes that have the highest ranks. These selected genes may be helpful to identify tumor-specific molecular markers. Moreover, our method considers the correlation among all genes and uses all genes to train the classifier, which is different from univariate feature selection. Thus, the selected marker genes may be more biologically meaningful.

Although LDA performs well in many applications, we cannot directly use LDA for cancer classification because of the *small sample size problem* [16], i.e. the number of samples is smaller than the dimensionality of samples. Recall that the rank of $\mathbf{S}_w$ is less than $n - c$ [16], where n is the number of samples and c is the number of classes. So, $\mathbf{S}_w$ would be singular (and thus LDA cannot be applied) if the number of samples minus the number of classes is smaller than the dimensionality of samples. This situation always happens in cancer classification since the data contains several thousand genes but only a few dozen samples. To overcome the small sample size problem, we develop a generalized linear discriminant analysis (GLDA) that works even when $\mathbf{S}_w$ is singular by taking advantage of some special properties of the scatter matrices as shown in the next section.

3 Generalized Linear Discriminant Analysis

In this section, we present the generalized linear discriminant analysis. The first subsection reviews some necessary details of the conventional LDA method. In this subsection, we will also briefly review previous work on solving the small sample size problem with LDA. In the second subsection, we present some properties of scatter matrices and our method to maximize Fisher's criterion, which works even when $\mathbf{S}_w$ is singular. Finally, we show a fast algorithm for GLDA in the last subsection.

3.1 Conventional Linear Discriminant Analysis

In order to find a $\mathbf{W}$ maximizing (1), we take the derivative of (1) with respect to $\mathbf{W}$ [16],

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{W}}\mathrm{tr}((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W}))\\&= -2\mathbf{S_w}\mathbf{W}(\mathbf{W}^T\mathbf{S_w}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})(\mathbf{W}^T\mathbf{S_w}\mathbf{W})^{-1}\\&\quad + 2\mathbf{S}_b\mathbf{W}(\mathbf{W}^T\mathbf{S_w}\mathbf{W})^{-1} \qquad (3)\end{aligned}$$

Equating (3) to zero, an optimal $\mathbf{W}$ must satisfy

$$\mathbf{S}_b\mathbf{W} = \mathbf{S}_w\mathbf{W}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W}) \qquad (4)$$

If $\mathbf{S}_w$ is nonsingular, we can multiply its inverse to both sides of Equation (4) and obtain the conventional LDA through simultaneous diagonalization of $\mathbf{S}_w$ and $\mathbf{S}_b$ [16]. When $\mathbf{S}_w$ is singular due to the small sample size problem, however, this procedure cannot be applied.

In recent years, many researchers have noticed this problem and tried to overcome the computational difficulty with LDA. A simple and direct attempt is to replace the inverse $\mathbf{S}_w^{-1}$ with the pseudo-inverse $\mathbf{S}_w^{+}$ [41]. However, it does not guarantee that Fisher's criterion is still optimized by the largest eigenvectors of $\mathbf{S}_w^{+}\mathbf{S}_b$. Another approach is to first reduce the dimensionality with some other feature selection/extraction method and then apply LDA on the dimensionality-reduced data. For instance, Belhumeur *et al.* proposed the Fisherface (also called PCA+LDA) method which first employs PCA to reduce the dimensionality of the feature space to $n-c$, and then applies the standard LDA to reduce the dimensionality to $c-1$ [3]. Note that this method is sub-optimal because PCA has to keep $n - 1$ principal components in order not to lose information. However, the first step of PCA+LDA keeps only $n - c$ principal components. Such a setting will lose too much information if the number of classes is large.

To handle the singularity problem, it is also popular to add a singular value perturbation to $\mathbf{S}_w$ to make it nonsingular [22]. A similar but more systematic method is regularized discriminant analysis (RDA) [15]. In RDA, one tries to obtain more reliable estimates of the eigenvalues by correcting the eigenvalue distortion in the sample covariance matrix with a ridge-type regularization. Besides, RDA is also a compromise between LDA and QDA (quadratic discriminant analysis), which allows one to shrink the separate covariances of QDA towards a common covariance as in LDA. Penalized discriminant analysis (PDA) is another regularized version of LDA [20, 21]. The goals of PDA are not only to overcome the small sample size problem but also to smooth the coefficients of discriminant vectors for better interpretation. In PDA, $\mathbf{S}_w$ is replaced with $\mathbf{S}_w + \lambda\Omega$ and then LDA proceeds as usual, where Ω is a symmetric and nonnegative definite penalty matrix. The choice of Ω depends on the problem. If the data are log-spectra or images, Ω is defined in such a way so as to force nearby components of discriminant vectors to be similar. The main problem with RDA and PDA is that they do not scale well. In applications such as face recognition and cancer classification with gene expression profiling, the dimensionality of covariance matrices are often more than ten thousand. It is not practical for RDA and PDA to process such large covariance matrices, especially when the computing platform is made of PCs.

Recently, several methods that play with the null space of $\mathbf{S}_w$ have been widely investigated. A well-known null subspace method is the LDA+PCA method [8]. When $\mathbf{S}_w$ is of full rank, the LDA+PCA method just calculates the largest eigenvectors of $\mathbf{S}_t^{-1}\mathbf{S}_b$ to form the transformation matrix. Otherwise, a two-stage procedure is employed. First, the data are transformed into the null space $\mathcal{V}_0$ of $\mathbf{S}_w$. Then, it tries to maximize the between-class scatter in $\mathcal{V}_0$, which is accomplished by performing PCA on the between-class scatter matrix in $\mathcal{V}_0$. Although this method solves the small sample size problem, it could be sub-optimal because it maximizes the between-class scatter in the null space of $\mathbf{S}_w$ instead of the original input space. For example, the performance of the LDA+PCA method drops significantly when $n - c$ is close to the dimensionality D. The reason is that the dimensionality of the null space $\mathcal{V}_0$ is very small in this situation, and too much information is lost when we try to extract the discriminant vectors in $\mathcal{V}_0$. LDA+PCA also needs to calculate the rank of $\mathbf{S}_w$, which is an ill-defined operation due to floating-point imprecision. Another problem with LDA+PCA is that the computational complexity of determining the null space of $\mathbf{S}_w$ is very high. In [23], a more efficient null subspace method was proposed, which has the same accuracy as LDA+PCA. This method first removes the null space of $\mathbf{S}_t$, which has been proven to be the common null space of $\mathbf{S}_b$ and $\mathbf{S}_w$, and useless for discrimination. Then, LDA+PCA is performed in the lower-dimensional projected space. Direct LDA is another null space method that discards the null space of $\mathbf{S}_b$ [46]. This is achieved by diagonalizing first $\mathbf{S}_b$ and then diagonalizing $\mathbf{S}_w$, which is in the reverse order of conventional simultaneous diagonalization procedure. In Direct LDA, one may also employ $\mathbf{S}_t$ instead of $\mathbf{S}_w$. In this way, Direct LDA is actually equivalent to the PCA+LDA [46]. Therefore, Direct LDA may be regarded as a "unified PCA+LDA" since there is no separate PCA step. Recently, we proposed a new feature extraction criterion, the maximum margin criterion (MMC), to avoid the small sample size problem [27]. A feature extractor based on MMC tries to maximize $\mathbf{S}_b - \mathbf{S}_w$ after the dimensionality reduction. From a geometric standpoint, MMC maximizes the (average) margin between classes. In what follows, a direct improvement of LDA is proposed to overcome the small sample size problem.

3.2 The Proposed Method

It is known that Fisher's criterion may also be written as

$$J(\mathbf{W}) = \text{tr}\left(\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_t\mathbf{W}}\right) \tag{5}$$

which is exactly equivalent to Equation (1) when $\mathbf{S}_w$ is nonsingular [16]. However, there is an important difference between Equations (1) and (5) when $\mathbf{S}_t$ is singular (so is $\mathbf{S}_w$). It is well known that the null space of $\mathbf{S}_w$ contains the important discriminatory information. As pointed out in [15], the discriminant function is heavily weighted by the smallest eigenvalues of $\mathbf{S}_w$ and the directions associated with their eigenvectors. When the small sample size problem occurs, these smallest eigenvalues are estimated to be 0. That is, the corresponding eigenvectors are in the null space of $\mathbf{S}_w$. Besides, it was also experimentally shown that the null space of $\mathbf{S}_w$ is crucial for discriminant analysis [8]. Unfortunately, we cannot take the derivative of Equation (1) in the null space of $\mathbf{S}_w$ to find the optimal solution because the vectors in the null space of $\mathbf{S}_w$ are the singular points of Equation (1). In contrast, the null space of $\mathbf{S}_t$ is a subspace of the null space of $\mathbf{S}_b$, which is useless for extracting the discriminatory information [23]. Thus, we can safely take the derivative of Equation (5) out of the null space of $\mathbf{S}_t$ in order to find the optimal solution. Therefore, we will use Equation (5) as the optimization criterion in the rest of paper.

In our proposed method and associated lemmas, the Moore-Penrose inverse is used, which is defined as:

Definition 1 (Moore-Penrose Inverse) *A matrix* $\mathbf{A}^+$ *satisfying the following conditions is unique and is called the Moore-Penrose inverse of* $\mathbf{A}$*:*

$$\begin{aligned} \mathbf{A}\mathbf{A}^+\mathbf{A} &= \mathbf{A} & \mathbf{A}^+\mathbf{A}\mathbf{A}^+ &= \mathbf{A}^+ \\ \left(\mathbf{A}^+\mathbf{A}\right)^T &= \mathbf{A}^+\mathbf{A} & \left(\mathbf{A}\mathbf{A}^+\right)^T &= \mathbf{A}\mathbf{A}^+ \end{aligned}$$

One may also define the matrix 1-inverse $\mathbf{A}^-$ by requiring only the first condition. However, such an inverse is not unique in general. Although most of our results also hold for the matrix 1-inverse, we confine ourselves to the Moore-Penrose inverse in this paper for uniquity.

The following lemmas will be needed in the proof of the main theorem.

Lemma 2 $\mathbf{S}_t\mathbf{S}_t^+(\mathbf{x} - \mathbf{m}) = \mathbf{x} - \mathbf{m}$

Proof. First, we can prove

$$E[(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)(\mathbf{x} - \mathbf{m})] = (\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)E[\mathbf{x} - \mathbf{m}] = 0$$

and

$$\begin{aligned} &\text{cov}((\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)(\mathbf{x} - \mathbf{m})) \\ &= (\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)\text{cov}(\mathbf{x} - \mathbf{m})(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)^T \\ &= (\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)\mathbf{S}_t(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)^T \\ &= (\mathbf{S}_t - \mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_t)(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)^T \\ &= (\mathbf{S}_t - \mathbf{S}_t)(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)^T = 0 \end{aligned}$$

So,

$$(\mathbf{I} - \mathbf{S}_t\mathbf{S}_t^+)(\mathbf{x} - \mathbf{m}) = 0$$

i.e.,

$$\mathbf{S}_t\mathbf{S}_t^+(\mathbf{x} - \mathbf{m}) = \mathbf{x} - \mathbf{m}$$

■

Lemma 2 means that any centered sample $\mathbf{x} - \mathbf{m}$ is the eigenvector of $\mathbf{S}_t\mathbf{S}_t^+$ corresponding to eigenvalue 1. In this sense, it is natural that both $\mathbf{S}_b$ and $\mathbf{S}_w$ are constituted by the eigenvectors of $\mathbf{S}_t\mathbf{S}_t^+$ because the column vectors of both $\mathbf{S}_b$ and $\mathbf{S}_w$ are just the linear combinations of centered samples. These will be shown in Lemmas 3 and 4 below.

Define

$$\mathbf{M} = [\sqrt{p_1}(\mathbf{m}_1 - \mathbf{m}), \ldots, \sqrt{p_c}(\mathbf{m}_c - \mathbf{m})]$$

where c is the number of classes; $\mathbf{m}_i$ and p_i are the mean vector and *a priori* probability of class i, respectively; and $\mathbf{m}$ is the overall mean vector. So,

$$\mathbf{S}_b = \mathbf{M}\mathbf{M}^T$$

and

$$\mathbf{M} = [E[\sqrt{p_1}(\mathbf{x} - \mathbf{m})|\mathcal{C}_1], \ldots, E[\sqrt{p_c}(\mathbf{x} - \mathbf{m})|\mathcal{C}_c]]$$

Lemma 3 $\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_b = \mathbf{S}_b$

Proof. By Lemma 2, we have

$$\mathbf{S}_t\mathbf{S}_t^+(\mathbf{x} - \mathbf{m}) = \mathbf{x} - \mathbf{m}$$

Obviously,

$$\mathbf{S}_t\mathbf{S}_t^+\sqrt{p_i}(\mathbf{x} - \mathbf{m}) = \sqrt{p_i}(\mathbf{x} - \mathbf{m})$$

and

$$\mathbf{S}_t\mathbf{S}_t^+E[\sqrt{p_i}(\mathbf{x} - \mathbf{m})|\mathcal{C}_i] = E[\sqrt{p_i}(\mathbf{x} - \mathbf{m})|\mathcal{C}_i]$$

for $i = 1, \ldots, c$. Thus,

$$\mathbf{S}_t\mathbf{S}_t^+\mathbf{M} = \mathbf{M}$$
$$\mathbf{S}_t\mathbf{S}_t^+\mathbf{M}\mathbf{M}^T = \mathbf{M}\mathbf{M}^T$$

i.e.,

$$\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_b = \mathbf{S}_b$$

■

Using Lemma 3, it is straightforward to prove a similar lemma for $\mathbf{S}_w$

Lemma 4 $\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_w = \mathbf{S}_w$

From Lemmas 3 and 4, we also have $\mathbf{S}_b\mathbf{S}_t^+\mathbf{S}_t = \mathbf{S}_b$ and $\mathbf{S}_w\mathbf{S}_t^+\mathbf{S}_t = \mathbf{S}_w$ since $\mathbf{S}_b$, $\mathbf{S}_w$, and $\mathbf{S}_t$ are symmetric.

The following theorem is the foundation of the GLDA method. That is, Fisher's criterion (Equation (5)) is maximized by the largest eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$. When $\mathbf{S}_w$ is nonsingular, $\mathbf{S}_t$ is also nonsingular and $\mathbf{S}_t^+$ is equal to $\mathbf{S}_t^{-1}$. Thus, GLDA coincides with the conventional LDA when $\mathbf{S}_w$ is nonsingular. Note that our method is very different from the naive method that simply replaces the inverse $\mathbf{S}_w^{-1}$ with the pseudo-inverse $\mathbf{S}_w^+$, which has no rigorous mathematical support. In fact, we know that Equation (1) is not valid when $\mathbf{S}_w$ is singular. It is meaningless to just use the pseudo-inverse $\mathbf{S}_w^+$ in this situation. On the other hand, we carefully analyze the properties of scatter matrices and if the null spaces of scatter matrices contain the discriminatory information. Our method is supported by the rigorous proofs (see below) and thus mathematically well-founded. Loosely speaking, GLDA can be regarded as a special case of PDA in the sense that $\Omega = \mathbf{S}_b$ and $\lambda = 1$. Note that Ω should be chosen intelligently in PDA so that $\mathbf{S}_t + \lambda\Omega$ is invertible [20]. However, $\mathbf{S}_t + \mathbf{S}_b$ is usually singular in GLDA. Different from the general PDA, GLDA has a closed-form solution. Later, we will also develop a fast algorithm of GLDA to efficiently handle the high dimensionality of data. In contrast, PDA encounters the computational difficulties when dealing with very high dimensional data.

Theorem 5 *Fisher's criterion (Equation* (5)*) is maximized by the largest eigenvectors of* $\mathbf{S}_t^+\mathbf{S}_b$.

Proof. Similar to the procedure for obtaining (4), we obtain the following equation by taking the derivative of (5) with respect to $\mathbf{W}$ and equating it to zero.

$$\mathbf{S}_b\mathbf{W} = \mathbf{S}_t\mathbf{W}(\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W}) \tag{6}$$

Note that we have to restrict the domain of $\mathbf{W}$ to be outside the null space of $\mathbf{S}_t$ in order to take the derivative. As we mentioned before, the null space of $\mathbf{S}_t$ is a subspace of the null space of $\mathbf{S}_b$, which does not contain any discriminatory information [23]. Thus, such a restriction does not limit the discriminant capacity of the method.

Since the null space of $\mathbf{S}_t$ is a subspace of the null space of $\mathbf{S}_b$ [23], we can simultaneously diagonalize two symmetric matrices $\mathbf{W}^T\mathbf{S}_b\mathbf{W}$ and $\mathbf{W}^T\mathbf{S}_t\mathbf{W}$ to $\mathbf{\Lambda}$ =diag$[\lambda_1, \ldots, \lambda_d]$ and $\mathbf{I}$ [16]:

$$\mathbf{P}^T(\mathbf{W}^T\mathbf{S}_b\mathbf{W})\mathbf{P} = \mathbf{\Lambda} \tag{7}$$
$$\mathbf{P}^T(\mathbf{W}^T\mathbf{S}_t\mathbf{W})\mathbf{P} = \mathbf{I} \tag{8}$$

where $\mathbf{P}$ is a $d \times d$ nonsingular matrix and d is less than or equal to the rank of $\mathbf{S}_t$. Besides, $\mathbf{\Lambda} \geq 0$ because $\mathbf{S}_b$ is positive semidefinite [16]. So, we have

$$\mathbf{W}^T\mathbf{S}_b\mathbf{W} = (\mathbf{P}^{-1})^T\mathbf{\Lambda}\mathbf{P}^{-1} \tag{9}$$
$$\mathbf{W}^T\mathbf{S}_t\mathbf{W} = (\mathbf{P}^{-1})^T\mathbf{P}^{-1} \tag{10}$$

Using Equation (9) and (10), we can simplify the right hand side of Equation (6) as follows:

$$\begin{aligned}&\mathbf{S}_t\mathbf{W}(\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})\\ &= \mathbf{S}_t\mathbf{W}[(\mathbf{P}^{-1})^T\mathbf{P}^{-1}]^{-1}[(\mathbf{P}^{-1})^T\mathbf{\Lambda}\mathbf{P}^{-1}]\\ &= \mathbf{S}_t\mathbf{W}\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}\end{aligned} \tag{11}$$

Then combining (6) and (11) leads to

$$\mathbf{S}_b\mathbf{W} = \mathbf{S}_t\mathbf{W}\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} \tag{12}$$
$$\mathbf{S}_b\mathbf{W}\mathbf{P} = \mathbf{S}_t\mathbf{W}\mathbf{P}\mathbf{\Lambda} \tag{13}$$

Plugging in $\mathbf{S}_b = \mathbf{S}_b\mathbf{S}_t^+\mathbf{S}_t$, we obtain

$$\mathbf{S}_b\mathbf{S}_t^+\mathbf{S}_t\mathbf{W}\mathbf{P} = \mathbf{S}_t\mathbf{W}\mathbf{P}\mathbf{\Lambda} \tag{14}$$

Denoting $\mathbf{K} = \mathbf{S}_t\mathbf{W}\mathbf{P}$, (14) can be expressed as

$$\mathbf{S}_b\mathbf{S}_t^+\mathbf{K} = \mathbf{K}\mathbf{\Lambda} \tag{15}$$

which means that the column vectors of $\mathbf{K}$ are the eigenvectors of $\mathbf{S}_b\mathbf{S}_t^+$ and $\lambda_i, i = 1, \ldots, d$ are the corresponding eigenvalues.

Because Fisher's criterion is invariant under any nonsingular linear transformation [16], we also have

$$\begin{aligned} &J(\mathbf{W}) = J(\mathbf{W}\mathbf{P}) \\ &= \text{tr}((\mathbf{P}^T\mathbf{W}^T\mathbf{S}_t\mathbf{W}\mathbf{P})^{-1}(\mathbf{P}^T\mathbf{W}^T\mathbf{S}_b\mathbf{W}\mathbf{P})) \\ &= \text{tr}(\mathbf{\Lambda}) = \lambda_1 + \cdots + \lambda_d \end{aligned}$$

Hence, in order to maximize the objective function, we have to choose the column vectors of $\mathbf{K}$ as the d largest eigenvectors of $\mathbf{S}_b\mathbf{S}_t^+$. Since $\mathbf{K} = \mathbf{S}_t\mathbf{W}\mathbf{P}$, we have

$$\begin{aligned} &J(\mathbf{S}_t^+\mathbf{K}) = J(\mathbf{S}_t^+\mathbf{S}_t\mathbf{W}\mathbf{P}) = J(\mathbf{S}_t^+\mathbf{S}_t\mathbf{W}) \\ &= \text{tr}((\mathbf{W}^T\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_t\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_b\mathbf{S}_t^+\mathbf{S}_t\mathbf{W})) \\ &= \text{tr}((\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})) = J(\mathbf{W}) \end{aligned}$$

because $\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_b = \mathbf{S}_b$ and $\mathbf{S}_t\mathbf{S}_t^+\mathbf{S}_t = \mathbf{S}_t$. In other words, the optimal transformation $\mathbf{W}$ is

$$\mathbf{W} = \mathbf{S}_t^+\mathbf{K} \tag{16}$$

By Equation (15), we have

$$\mathbf{S}_t^+\mathbf{S}_b\mathbf{S}_t^+\mathbf{K} = \mathbf{S}_t^+\mathbf{K}\mathbf{\Lambda} \tag{17}$$

$$\mathbf{S}_t^+\mathbf{S}_b\mathbf{W} = \mathbf{W}\mathbf{\Lambda} \tag{18}$$

i.e., the column vectors of $\mathbf{W}$ are the eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$. ■

Since the rank of $\mathbf{S}_b$ is $c-1$, there are only $c-1$ eigenvectors to constitute the mapping matrix $\mathbf{W}$. In LDA/GLDA, we also require $\mathbf{W}$ to be orthonormal, which help preserve the shape of the distribution of the data.

3.3 A Fast Algorithm

Although the above GLDA solves the small sample size problem, we cannot apply it to cancer classification without developing a fast algorithm given the dimensionality of gene expression data. Note that both $\mathbf{S}_t$ and $\mathbf{S}_b$ have the dimensionality $D \times D$, where D is the number of genes. Since D is usually many thousands, it is very time and memory consuming to calculate $\mathbf{S}_t^+$ and the eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$ using common computational methods. In what follows, we will devise a fast algorithm to efficiently calculate $\mathbf{S}_t^+$ and the eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$ via singular value decomposition (SVD). SVD expresses a real $n \times m$ matrix $\mathbf{A}$ as a product $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$, where $\mathbf{\Lambda}^{\frac{1}{2}}$ is a diagonal matrix with decreasing non-negative entries, and $\mathbf{U}$ and $\mathbf{V}$ are $n \times \min(n, m)$ and $m \times \min(n, m)$ orthonormal column matrices [18]. The columns of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$, respectively, and the nonvanishing entries of $\mathbf{\Lambda}^{\frac{1}{2}}$ are the square roots of the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

```
ALGORITHM Generalized Linear Discriminant Analysis
Input: A gene expression dataset containing $n$ samples with corresponding labels from $c$ classes.
Output: The mapping matrix $\mathbf{W}$.
Method:
  Calculate $\mathbf{M} = [\sqrt{p_1}(\mathbf{m}_1 - \mathbf{m}), \ldots, \sqrt{p_c}(\mathbf{m}_c - \mathbf{m})]$ and $\mathbf{X} = \frac{1}{\sqrt{n}}[(\mathbf{x}_1 - \mathbf{m}), \ldots, (\mathbf{x}_n - \mathbf{m})]$.
  Perform the SVD $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$
  $\mathbf{S}_t^{+\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$
  Perform the SVD $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}^{\frac{1}{2}}\widetilde{\mathbf{V}}^T$
  $\mathbf{W} = \mathbf{S}_t^{+\frac{1}{2}}\widetilde{\mathbf{U}}$
```

Figure 1. A fast algorithm of generalized linear discriminant analysis.

Since $\mathbf{S}_t$ is symmetric, we can calculate its Moore-Penrose inverse through eigen decomposition. Suppose $\mathbf{S}_t = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, where the columns $\mathbf{u}_i$ of $\mathbf{U}$ are the mutually orthonormal eigenvectors of $\mathbf{S}_t$, and $\mathbf{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues λ_i. The Moore-Penrose inverse is $\mathbf{S}_t^+ = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T$. Note that $\mathbf{S}_t$ is estimated by $\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$ and can be expressed in the form $\mathbf{S}_t = \mathbf{X}\mathbf{X}^T$ with $\mathbf{X} = \frac{1}{\sqrt{n}}[(\mathbf{x}_1 - \mathbf{m}), \ldots, (\mathbf{x}_n - \mathbf{m})]$. Thus, we can obtain the eigenvalues λ_i and the corresponding orthonormal eigenvectors $\mathbf{u}_i$ of $\mathbf{S}_t$ through the SVD of $\mathbf{X}$. Note that dimensionality of $\mathbf{X}$ is $D \times n$, where n is the number of samples. Since n is only a few dozen and is much smaller than D in practice, the SVD of $\mathbf{X}$ is much faster than the eigen decomposition of $\mathbf{S}_t$.

To obtain the eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$, we follow an indirect approach. Let $\mathbf{S}_t^{+\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^T$. By Equation (18), we have

$$\mathbf{S}_t^{+\frac{1}{2}}\mathbf{S}_b\mathbf{S}_t^+\mathbf{K} = \mathbf{S}_t^{+\frac{1}{2}}\mathbf{K}\widetilde{\mathbf{\Lambda}}$$

$$\mathbf{S}_t^{+\frac{1}{2}}\mathbf{S}_b\mathbf{S}_t^{+\frac{1}{2}}(\mathbf{S}_t^{+\frac{1}{2}}\mathbf{K}) = (\mathbf{S}_t^{+\frac{1}{2}}\mathbf{K})\widetilde{\mathbf{\Lambda}}$$

i.e., the column vectors of $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{K}$ are the eigenvectors of $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{S}_b\mathbf{S}_t^{+\frac{1}{2}}$. Recall $\mathbf{S}_b = \mathbf{M}\mathbf{M}^T$ with $\mathbf{M} = [\sqrt{p_1}(\mathbf{m}_1 -$

$\mathbf{m}), \ldots, \sqrt{p_c}(\mathbf{m}_c - \mathbf{m})]$, we obtain $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{S}_b\mathbf{S}_t^{+\frac{1}{2}} = (\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M})(\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M})^T$. Thus, we can obtain $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{K}$ by the SVD of $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M}$. Since $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M}$ has the dimensionality $D \times c$ and the number of classes c is very small, it is very fast to compute the SVD of $\mathbf{S}_t^{+\frac{1}{2}}\mathbf{M}$. Finally, we get $\mathbf{W} = \mathbf{S}_t^{+\frac{1}{2}}(\mathbf{S}_t^{+\frac{1}{2}}\mathbf{K})$ by Equation (16).

A concise description of the algorithm is shown in Figure 1. Note that we may skip calculating $\mathbf{S}_t^{+\frac{1}{2}}$ in step 3. When $\mathbf{S}_t^{+\frac{1}{2}}$ is needed later, we can calculate it on the fly to save memory by arranging the order of matrix multiplications.

4 Experiments

In this section, we extensively compare our new method with other methods in the literature on many public cancer datasets. First, we compare GLDA with PDA [20, 21]. Then, we compare our method with a typical univariate feature selection method [11]. This also serves to compare our method with other widely used classifiers in cancer classification. We will also compare our method with GA/KNN, which is a multivariate feature selection methods [28, 29]. Finally, we compare our method with support vector machine with recursive feature elimination (RFE) [33].

4.1 Data

In the experiments, we test our new method on seven public datasets, Leukemia [19], Colon [2], Prostate [37], Lymphoma [1], SRBCT [26], Brain [32], and GCM [33]. The leukemia dataset comprises 47 acute lymphoblastic leukemia (ALL, 38 B-cell ALL and 9 T-cell ALL) and 25 actue myeloid leukemia (AML) samples with the expression levels of 3571 genes. The colon dataset has the expression levels of 2000 filtered genes in 40 tumor and 22 normal colon tissues. The prostate dataset contains the expression levels of 6033 filtered genes in 52 tumor and 50 normal prostate tissues. The lymphoma dataset contains 62 samples of 3 classes: 42 diffuse large B-cell lymphoma, 9 observations of follicular lymphoma, and 11 cases of chronic lymphocytic leukemia. The expression levels of 4026 genes in these samples are used in the experiments. The SRBCT data consists of 63 samples of four subclasses of small, round blue cell tumors of childhood (SRBCTs), which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). The number of genes is 2308. The brain dataset is the dataset A in [32] that has 42 samples of 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors (AT/RTs), 8 primitive neuroectodermal tumors (PNETs) and 4 normal cerebella. The brain dataset contains the expression of 5597 genes. The GCM dataset is a very complicated dataset, which is a collection of 14 primary human cancer classes with the expression levels of 16063 genes [33]. These 14 common tumor classes account for $\approx 80\%$ of new cancer diagnoses in the U.S. The GCM data consist of one training dataset of 144 primary tumor samples and one test dataset of 54 samples (46 primary and 8 metastatic). We combine the 144 samples of training dataset and 46 primary tumor samples of test dataset together and use the total 190 samples in the experiments. Before classification, the gene expression data are usually preprocessed in practice. In the experiments, we use the preprocessed data of the first six datasets by M. Dettling [9], which can be downloaded from http://stat.ethz.ch/∼dettling/bagboost.html. The preprocessing procedure includes base 10 log-transformation, normalization (with mean 0 and variance 1), and missing value imputation (by k-nearest neighbor) [9, 11]. Because the values in the GCM dataset are the raw average difference values (maybe negative) output from the Affymetrix software package, we do not perform the log-transformation and only normalize the values to mean 0 and variance 1. The properties of datasets are summarized in the top row of Table 1.

4.2 Results

The experimental procedure is as follows. For each dataset, we randomly split it into three parts in a class-proportional manner, of which two parts are used for training (both feature selection/extraction methods and classifiers) and the last part is kept for test. This procedure is repeated for 200 times and the averages and standard deviations of error rates are listed in Table 1. As shown in the table, GLDA performs well overall. To compare GLDA with other methods that try to solve the small sample size problem, we also perform PDA on the datasets because PDA has a nice mathematical foundation and shows good performance in some applications [20, 21]. We observe that PDA and GLDA achieve similar results on Leukemia, Colon, Lymphoma, and SRBCT. However, PDA could not be applied on the Prostate, Brain, and GCM because the datasets have very high dimensionalities and thus the available memory (1GB on our machine) is not sufficient for PDA.

In what follows, we compare our method with a univariate feature selection method. Because it was reported that various feature selection methods have the similar performance for cancer classification [30], it is sufficient to compare our method only with a typical feature selection method. In particular, we compare our method with the univariate feature selection method of Dudoit *et al.* [11], which employs a metric similar to LDA:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \qquad (19)$$

Table 1. Average classification error rates and standard deviations on seven public datasets based on 200 runs. In the table, c is the number of classes, n is the number of samples, D is the number of genes, and RandFor stands for random forests.

	Features	Leukemia	Colon	Prostate	Lymphoma	SRBCT	Brain	GCM
c		2	2	2	3	4	5	14
n		72	62	102	62	63	42	190
D		3571	2000	6033	4026	2308	5597	16063
GLDA	$c-1$	3.1 ± 2.8	14.5 ± 5.7	7.6 ± 3.7	0.05 ± 0.47	1.9 ± 2.6	16.6 ± 8.8	17.9 ± 3.8
PDA	$c-1$	3.3 ± 2.7	14.0 ± 5.7	N/A	0.17 ± 0.88	1.7 ± 2.4	N/A	N/A
RandFor	200	3.0 ± 3.2	14.6 ± 6.2	8.0 ± 4.5	0.76 ± 2.31	2.0 ± 2.8	23.6 ± 8.3	28.1 ± 4.5
SVM	200	2.0 ± 2.5	13.7 ± 6.4	8.6 ± 4.5	1.07 ± 2.26	2.3 ± 3.1	22.5 ± 9.7	32.8 ± 4.4
DLDA	200	2.8 ± 2.9	14.4 ± 6.6	15.5 ± 7.7	1.71 ± 2.57	2.1 ± 2.7	24.0 ± 9.7	31.6 ± 5.0
kNN	200	3.7 ± 3.1	18.2 ± 6.4	11.2 ± 4.6	1.12 ± 2.24	0.9 ± 1.9	23.0 ± 8.3	44.1 ± 4.6
RandFor	50	3.8 ± 3.6	14.8 ± 6.0	7.9 ± 4.8	2.29 ± 3.63	1.6 ± 2.9	25.3 ± 9.8	34.1 ± 4.6
SVM	50	3.2 ± 3.4	13.5 ± 5.8	8.4 ± 4.5	2.07 ± 3.29	1.8 ± 3.2	25.7 ± 10.1	39.3 ± 4.6
DLDA	50	2.8 ± 2.9	13.6 ± 6.4	12.3 ± 6.5	3.86 ± 3.92	1.8 ± 3.6	25.8 ± 10.5	38.2 ± 5.2
kNN	50	4.2 ± 3.4	19.3 ± 7.2	12.6 ± 4.8	2.12 ± 3.46	1.8 ± 2.6	26.6 ± 9.6	46.8 ± 5.8
RandFor	10	4.6 ± 3.8	15.9 ± 6.4	8.8 ± 4.3	3.93 ± 4.33	17.0 ± 10.3	39.2 ± 12.5	45.4 ± 5.0
SVM	10	3.5 ± 3.5	13.3 ± 5.7	8.2 ± 4.2	4.76 ± 5.01	19.0 ± 10.6	38.8 ± 12.1	49.8 ± 5.8
DLDA	10	3.2 ± 3.5	12.9 ± 6.0	11.0 ± 5.6	7.76 ± 5.85	17.4 ± 10.1	37.3 ± 14.0	49.6 ± 6.0
kNN	10	3.6 ± 3.5	18.6 ± 6.8	11.0 ± 4.7	5.00 ± 5.16	21.6 ± 12.0	43.9 ± 12.9	54.0 ± 4.7

where $\bar{x}_{\cdot j}$ denotes the average expression level of gene j across all samples, $\bar{x}_{kj}$ denotes the average expression level of gene j across samples belonging to class k, x_{ij} is the expression level of gene j in sample i, and y_i is the class label of sample i. The genes with the largest BSS/WSS ratios will be used for training. In the experiments, we try to select various numbers of top ranked genes. Due to the space limit, we report only the results with 10, 50, and 200 top ranked genes here. We choose this feature selection method for comparison because it has a similar standpoint as LDA. However, it selects genes independently but LDA fully considers the correlation among genes. Thus, this provides a good opportunity to investigate if our method improves the performance of classification compared with univariate feature selection. After feature selection, we use linear support vector machines (SVM) [43], random forests [7], k-nearest neighbor (kNN, $k = 1$ here) [13], and diagonal linear discriminant analysis (DLDA) [11] [1] for classification. Thus, the experiments also serve as a comparison between our method and the aforementioned classification methods. SVM and random forests are state-of-the-art machine learning methods and have proven very powerful in many applications. Although DLDA and kNN are very simple, many researcher have reported that they work very well for cancer classification. The weighted voting method of Golub *et al.* [19] is actually a minor variant of DLDA [11], and thus we do not include it in the experiments.

With 200 top ranked genes determined by the method of Dudoit *et al.*, all aforementioned methods achieve similar results as GLDA on Leukemia, Colon, Prostate, Lymphoma, and SRBCT. However, GLDA achieves much better accuracies than its competitors on the Brain and GCM, which are very hard instances because of their large numbers of classes and genes. Compared with other methods, our method reduces the error rate from about 22.5% to 16.6% on Brain and from about 28.1% to 17.9% on GCM. Besides, we observe that, one cannot meet the recommended sample per class ratio (i.e., 5 – 10) with 200 selected genes. Thus, the estimated error rates of the competing methods may be highly biased and the reliability of their results is low. In contrast, we can easily meet the recommended ratio after the dimension reduction by GLDA, which makes the estimated error rates more accurate and trustable. For a meaningful comparison (of trustable error rates), it is more suitable to compare the accuracy between GLDA and other classifiers with 10 top ranked genes. For such a comparison, GLDA is clearly better than all other methods on all datasets except for Colon. For Colon, we observe that all methods have roughly the same (high) error rates and the error rates do not change much with different numbers of genes. It has been reported that the colon dataset has a sample contamination problem [28], and may not be a suitable benchmark dataset.

Besides univariate feature selection, we would also like

[1] DLDA is actually the linear maximum likelihood discriminant rule that assumes the diagonal covariance matrix [11].

Table 2. The averages and standard deviations of the error rates of GA/KNN given 10, 50, or 200 top ranked genes.

Datasets	Runs	10	50	200
Prostate	100	7.3 ± 3.5	8.1 ± 4.1	8.6 ± 4.4
SRBCT	200	3.0 ± 3.4	1.3 ± 2.3	1.4 ± 2.3
Brain	200	31.5 ± 9.9	22.1 ± 8.2	21.0 ± 8.2
GCM	25	55.6 ± 5.0	41.7 ± 5.0	36.3 ± 3.6

to compare our method with the multivariate feature selection methods. Currently, two multivariate methods, gene-pair-ranking [6] and GA/KNN [28, 29], have been proposed in the literature. The gene-pair-ranking method gives each pair of genes a score reflecting how well the pair in combination distinguishes two classes. Because the gene-pair-ranking method does not show a significant superiority to other methods and is hard to extend to deal with general gene subsets due to time complexity [6], we compare our method only with GA/KNN below. Since most methods can achieve a high accuracy on Leukemia, and Lymphoma with only 10 top genes ranked by univariate methods and Colon has a sample contamination problem, we only show the comparison on the Prostate, SRBCT, Brain, and GCM datasets. The experimental procedure is the same as before and the results are summarized in Table 2. However, GA/KNN is very slow and needs a couple of weeks/months to complete 200 runs on Prostate/GCM datasets. As a result, we only perform GA/KNN 100 and 25 times on Prostate and GCM, respectively. As shown in Table 1 and 2, our method is better than GA/KNN overall although GA/KNN performs better than univariate feature selection methods. Note that GA/KNN achieves worse results with more genes on the Prostate dataset, which is different from the trends on other datasets. The reason is not clear. In principle, one may improve the performance of GA/KNN by replacing k-nearest neighbor with some advanced classifier such as SVM. Unfortunately, it will increase the running time significantly since these (advanced) classifiers have a much higher time complexity than kNN, which has no training procedure. In fact, GA/KNN is already very slow even with kNN. For example, it takes more than five days to complete 200 runs on Brain on an Athlon MP 2800+ machine. Our method, on the other hand, needs only several minutes. Recall that, the user of GA/KNN also has to determine many parameters, such as chromosome length, the number of chromosomes, termination metric, etc. In contrast, GLDA does not need any parameters. Given the results in Table 1 and 2, we also observe that the simple methods such as GLDA and GA/KNN that multivariately find marker genes can achieve better results than sophisticated classification methods (e.g. random forests and SVM) that combined with a univariate feature selection method, especially when the number of selected features is small. It indicates that the choice of feature selection/extraction methods may be more important than the choice of classifiers for cancer classification.

Table 3. The leave-one-out cross validation accuracy on the GCM training dataset. The left part is the results on genes selected by RFE that were obtained by Ramaswamy *et al.* [33]. The right part is the results on genes selected by GLDA. In the table, s is the number of selected genes per classifier.

RFE			GLDA		
s	kNN OVA	SVM OVA	Avg. Genes	kNN	GLDA
30	65.3%	70.8%	212	67.4%	78.5%
92	68.0%	72.2%	461	65.3%	81.9%
281	65.7%	73.4%	1132	66.0%	81.9%
1073	66.5%	74.1%	3972	66.7%	85.4%
3276	66.3%	74.7%	9821	67.4%	85.4%
6400	64.2%	75.5%	14512	67.4%	84.7%
All	N/A	78.0%	All	67.4%	84.7%

In [33], Ramaswamy *et al.* proposed recursive feature elimination (RFE) that uses SVMs for both classification and feature selection. Recall that LDA can also be used to select genes by treating the elements in the mapping matrix as the weights of genes. Here we would like to compare our method with RFE. Like RFE, we first train LDA on all features/genes. Then, we select a subset of genes and train LDA again on the selected genes. Ramaswamy *et al.* applied their method on the GCM data. Since this is a multiclass problem and SVM is a binary classifier, Ramaswamy *et al.* tried both one-versus-all (OVA) and all-pairs (AP) output coding schemes. We list their leave-one-out cross validation results of OVA on 144 training samples in Table 3. The results based on AP are not listed here since they are worse than those based on OVA. For the OVA coding scheme, one need train c binary SVMs, where c is the number of classes ($c = 14$ in the discussion below). Each SVM uses its own selected genes. Thus, the total number of selected genes is $s \times c$, where s is the number of genes per classifier and listed in the first column of Table 3. Of course, there may exist some overlap among the marker genes of different SVMs. Because our GLDA can solve multiclass problems directly, we use a different experimental setting. After training GLDA, we choose the top $s \times (c-1)$ ranked genes because there are only $c-1$ eigenvectors of $\mathbf{S}_t^+\mathbf{S}_b$. Here, s is the same as that in RFE for reasonable comparison. Due to the overlap among top ranked genes of each eigenvector, the total number of genes is less than $s \times (c-1)$. The actual average numbers of selected genes are listed in the fourth column of Table 3. We observe that the overlap rate is very high. For instance, the overlap rate is $1 - 461/(92 \times 13) = 61.5\%$ for $s = 92$, which indicates that most eigenvectors give high weights to these 461 genes. Note that our method selects the genes in one step, not recursively. After the selection, we perform kNN and GLDA on the selected genes. Compared with RFE, GLDA performs clearly better as shown in Table 3. The accuracy 85.4% is the highest reported accuracy on GCM as far as we know. It is also interesting that the accuracy of SVM decreases as the number of marker genes decreases. Our method, however, does not show such a pattern. When the number of genes decreases, GLDA may achieve a better accuracy because the noise might be reduced. Furthermore, the accuracy decreases when the number of genes becomes too small because a lot of information may be lost. Finally, we observe that kNN performs slightly better with our feature selection method than with RFE.

5 Conclusion

Gene expression profiling has great potential for accurate cancer diagnosis. It also brings machine learning researchers two challenges, the curse of dimensionality and the small sample size problem. In this paper, we have presented a novel method to solve these two problems. Our extensive experiments on seven public datasets demonstrate that the method is able to classify tumors robustly with a high accuracy. Besides cancer classification, our work on generalized linear discriminant analysis may also find applications in other areas where the small sample size problem and the curse of dimensionality arise, such as image recognition and web document classification.

Acknowledgement

We thank Dr. M. Dettling for providing the six normalized cancer datasets and Dr. L. Li for providing the source code of GA/KNN. This work was partially supported by NSF grants CCR-0309902 and National Key Project for Basic Research (973) grant 2002CB512801.

References

[1] A. A. Alizadeh, E. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. *Proceedings of the National Academy of Sciences of USA*, 96(12):6745–6750, 1999.

[3] 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*, 19(7):711–720, 1997.

[4] R. E. Bellman. *Adaptive Control Precesses: A Guided Tour*. Princeton University Press, Princeton, NJ, 1961.

[5] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *Proceedings of the 4th Annual International Conference on Research in Computational Molecular Biology*, pages 54–64, 2000.

[6] T. H. Bø and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):1–11, 2002.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] 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.

[9] M. Dettling. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593, 2004.

[10] M. Dettling and P. Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–1069, 2003.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[12] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annual of Eugenics*, 7:179–188, 1936.

[13] E. Fix and J. Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[14] D. H. Foley. Considerations of sample and feature size. *IEEE Transactions on Information Theory*, 18(5):618–626, 1972.

[15] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

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

[17] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[18] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[19] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 536:531–537, 1999.

[20] T. Hastie and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102, 1995.

[21] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270, 1994.

[22] Z.-Q. Hong and J.-Y. Yang. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, 24(4):317–324, 1991.

[23] R. Huang, Q. Liu, H. Lu, and S. Ma. Solving the small sample size problem of LDA. In *Proceedings of 16th International Conference on Pattern Recognition*, volume 3, pages 29–32, 2002.

[24] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P. Krishnaiah and L. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. Amsterdam, North Holland, 1982.

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

[26] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

[27] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Advances in Neural Information Processing Systems 16*, pages 97–104, 2003.

[28] L. Li, T. A. Darden, C. R. Weinberg, A. J. Levine, and L. G. Pedersen. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinatorial Chemistry & High Throughput Screening*, 4(8):727–739, 2001.

[29] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12):1131–1142, 2001.

[30] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[31] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. AI Memo 1677, MIT CBCL, Cambridge, MA, 1998.

[32] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. S. G, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[33] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science*, 98(26):15149–15154, 2001.

[34] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10:159–203, 1948.

[35] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991.

[36] S. J. Raudys and V. Pikelis. On dimensionality, sample size, classification error, and complexity of classification algorithms in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:243–251, 1980.

[37] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[38] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class prediction and discovery using gene expression data. In *Proceedings of the 4th Annual International Conference on Research in Computational Molecular Biology*, pages 263–272, 2000.

[39] R. L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[40] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–1579, 2003.

[41] Q. Tian, M. Barbero, Z.-H. Gu, and S. H. Lee. Image classification by the foley-sammon transform. *Optical Engineering*, 25(7):834–840, 1986.

[42] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[43] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[44] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In T. K. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2000.

[45] C.-H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17(Supplement 1):316–322, 2001.

[46] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data — with application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

[47] H. Zhang, C.-Y. Yu, and B. Singer. Cell and tumor classification using gene expression data: construction of forests. *Proceedings of the National Academy of Sciences of USA*, 100(7):4168–4172, 2003.