

Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data

Jieping Ye, Tao Li, Tao Xiong, and Ravi Janardan

Abstract—The classification of tissue samples based on gene expression data is an important problem in medical diagnosis of diseases such as cancer. In gene expression data, the number of genes is usually very high (in the thousands) compared to the number of data samples (in the tens or low hundreds); that is, the data dimension is large compared to the number of data points (such data is said to be undersampled). To cope with performance and accuracy problems associated with high dimensionality, it is commonplace to apply a preprocessing step that transforms the data to a space of significantly lower dimension with limited loss of the information present in the original data. Linear Discriminant Analysis (LDA) is a well-known technique for dimension reduction and feature extraction, but it is not applicable for undersampled data due to singularity problems associated with the matrices in the underlying representation. This paper presents a dimension reduction and feature extraction scheme, called Uncorrelated Linear Discriminant Analysis (ULDA), for undersampled problems and illustrates its utility on gene expression data. ULDA employs the Generalized Singular Value Decomposition method to handle undersampled data and the features that it produces in the transformed space are uncorrelated, which makes it attractive for gene expression data. The properties of ULDA are established rigorously and extensive experimental results on gene expression data are presented to illustrate its effectiveness in classifying tissue samples. These results provide a comparative study of various state-of-the-art classification methods on well-known gene expression data sets.

Index Terms—Microarray data analysis, discriminant analysis, generalized singular value decomposition, classification.



1 INTRODUCTION

DNA microarrays, pioneered in [7], [10], are novel technologies that are designed to measure gene expression levels of tens of thousands of genes in a single experiment. The ability to measure gene expression levels for a very large number of genes, covering the entire genome in the case of some small organisms, raises the issue of characterizing cells in terms of gene expression levels, that is, using gene expression data to determine the state and functions of the cells. The most fundamental of the characterization problems is that of identifying a set of genes and their expression patterns that either characterize a certain cell state or predict a certain cell state in the future. A key step in this direction is the development of tools for classifying tissue samples according to their gene expression levels. Specifically, given a collection of gene expression data, grouped into classes according to, say, disease type, the goal is to decide which class a new tissue sample likely belongs to. Further discussion on this can be found in [1], [3], [5], [9], [11], [13], [15], [22], [25], [35].

Gene expression data possess characteristics that makes the task of classification quite challenging. Expression data usually contains a large number of genes (typically, in the thousands) and a small number of samples (typically, in the

tens or low hundreds). In machine learning terminology, these data sets are said to have high dimension and small sample size (i.e., undersampled data sets). Many methods have been proposed in the past to reduce the data dimensionality by selecting only a subset of the most relevant genes. In expression data, many gene groups interact closely. Gene interactions are important biologically and may contribute to class distinctions. Ignoring them is not desirable. Most gene selection methods, however, rely on rank-based schemes [8] and tend to ignore correlations between genes [23].

Linear Discriminant Analysis [12], [16] is a well-known scheme for feature extraction and dimension reduction. It has been used widely in many applications such as face recognition [4], text classification [17], [32], microarray data classification [8], etc. Given a set of high-dimensional data grouped into classes, classical LDA aims to find an optimal transformation that maps the data into a lower-dimensional space (while preserving the class structure) that minimizes the within-class distance and simultaneously maximizes the between-class distance, thus achieving maximum discrimination. The optimal transformation is readily computed by applying the eigen-decomposition on the scatter matrices. An intrinsic limitation of classical LDA is that its objective function requires that at least one of the scatter matrices be nonsingular. However, for many applications, such as microarray data classification, face recognition, and text classification, all scatter matrices in question can be singular since the data are from a very high-dimensional space and, in general, the dimension exceeds the sample size. This is known as the *undersampled* or *singularity* problem. More details can be found in [20], [32].

This paper proposes ULDA, a feature extraction and dimension reduction algorithm based on discriminant analysis. This method has several advantages. First, the

• J. Ye and R. Janardan are with the Department of Computer Science and Engineering, University of Minnesota, Twin Cities, 4-192 EE/CSci Bldg., 200 Union Street S.E., Minneapolis, MN 55455. E-mail: {jjeeping, janardan}@cs.umn.edu.

• T. Li is with the School of Computer Science, Florida International University, ECS 318, Miami, FL 33199. E-mail: taoli@cs.fiu.edu.

• T. Xiong is with the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities, 200 Union Street S.E., Minneapolis, MN 55455. E-mail: txiong@ece.umn.edu.

Manuscript received 8 Sept. 2004; revised 1 Dec. 2004; accepted 4 Dec. 2004. For information on obtaining reprints of this article, please send e-mail to: tccb@computer.org, and reference IEEECS Log Number TCBB-0026-0204.

transformed features obtained via ULDA are linear combinations of the original genes. In other words, ULDA takes the relationship between different genes into consideration. Second, we show theoretically that the features in the transformed space are uncorrelated, thus ensuring minimum redundancy among the features in the reduced space. Third, as we know, classical discriminant analysis fails if the within-class scatter matrix is singular [12]. For gene expression data, the within-class scatter matrix is usually singular owing to the large number of genes and the small size of observations. ULDA utilizes Generalized Singular Value Decomposition [14] to overcome this undersampling problems. Recent extensions of ULDA, as well as more detailed theoretical analysis can be found in [30], [31].

The rest of the paper is organized as follows: Section 2 provides background on related work. Classical Linear Discriminant Analysis is introduced in Section 3. Section 4 proposes ULDA for the special case, where the scatter matrix is nonsingular. Section 5 presents ULDA for the general case, which deals with the undersampled problems. Experimental results are presented in Section 6. Finally, Section 7 offers conclusions and further discussion.

2 RELATED WORK

There has been extensive research in tissue classification based on gene expression data. These methods include K-Nearest Neighbor (KNN), decision tree, naive Bayesian, bagging, boosting, discriminant analysis, Diagonal Linear Discriminant Analysis, Nearest Shrunken Centroid classifier, Support Vector Machines, and Multicategory Support Vector Machines (see, for example, [8], [21], [28]). In the following, we give a brief overview of four methods used in our empirical study.

K-Nearest Neighbor (KNN): KNN is a nonparametric classifier and theoretical studies have shown that its error is asymptotically at most twice that of the Bayesian error rate. KNN finds the K-nearest neighbors among training samples and uses the categories of the K neighbors to determine the category of the test sample. The parameter K for the number of neighbors is usually quite small (usually $K = 1$) due to the small size of the data sets. KNN was previously studied for tumor classification in [8].

Diagonal Linear Discriminant Analysis (DLDA): Diagonal Linear Discriminant Analysis [8] is a simplification of classical LDA, which applies the common diagonal covariance matrix to all classes. It is computationally more efficient than other LDA-based algorithms. Interestingly, the “weighted voting scheme” for binary classification proposed in [15] can be shown to be a variant of DLDA.

Nearest Shrunken Centroid (NSC): Nearest Shrunken Centroid classification [28] is an extension of standard nearest centroid classification. It “shrinks” each of the class centroids toward the overall centroid by an amount called the *threshold*. This shrinking consists of moving the centroid toward zero by the threshold amount, and setting it equal to zero if it hits zero. After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids. This shrinking has two advantages: 1) it can make the classifier more accurate by reducing the noise, and 2) it does automatic gene selection. However, the choice of the threshold using cross-validation for a range

TABLE 1
Summary of Notations Used in the Paper

Notation	Description
A	gene expression data matrix
n	number of data points
p	number of dimensions
k	number of classes
S_i	covariance matrix of the i -th class
S_b	between-class scatter matrix
S_w	within-class scatter matrix
S_t	total scatter matrix
G	transformation matrix
m_i	centroid of the i -th class
m	global centroid of the training set

of threshold values can be expensive, especially for large and high-dimensional data.

Support Vector Machines (SVM): Support Vector Machines (SVMs) [6], [29] have shown superb performance in binary classification tasks. Basically, Support Vector Machine computes a hyperplane that separates two classes of data with so-called maximum margin. To extend SVM to multicategory classification tasks, several methods have been proposed including those solving multiclass SVM in one step, and those based on binary classifications, such as one-against-all, one-against-one, and directed acyclic graph SVM (DAGSVM). Hsu and Lin did extensive experiments in [18] showing that the one-against-one and DAG methods are more suitable for practical use than the other methods. We therefore use the one-against-one scheme in our comparative studies.

3 CLASSICAL LINEAR DISCRIMINANT ANALYSIS

The notations used in the remainder of this paper are listed in Table 1.

Given a gene expression data matrix $A = (a_{ij}) \in \mathbb{R}^{p \times n}$, where each column corresponds to a sample and each row corresponds to a particular gene, we consider finding a linear transformation $G \in \mathbb{R}^{p \times \ell}$ ($\ell < p$) that maps each column a_i of A , for $1 \leq i \leq n$, in the p -dimensional space to z_i in the ℓ -dimensional space. That is,

$$G : a_i \in \mathbb{R}^p \rightarrow z_i = G^T \cdot a_i \in \mathbb{R}^\ell. \quad (1)$$

The resulting data matrix $A^L = G^T \cdot A \in \mathbb{R}^{\ell \times n}$ contains ℓ rows, i.e., there are ℓ features for each sample in the reduced (transformed) space. The features in the reduced space are linear combinations of the features in the original high-dimensional space, where the coefficients

of the linear combinations depend on the transformation matrix G .

A common way to compute the transformation matrix G is through classical Linear Discriminant Analysis (LDA). Classical LDA aims to compute the optimal transformation matrix G such that the class structure is preserved. More details are given below.

Assume that there are k classes in the data set. Suppose that m_i , S_i , P_i are the centroid, covariance matrix, and a prior probability of the i th class, respectively, and that m is the global centroid. Then, the between-class scatter matrix S_b , the within-class scatter matrix S_w , and the total scatter matrix S_t are defined as follows [12]:

$$S_b = \sum_{i=1}^k P_i (m_i - m)(m_i - m)^T, \quad (2)$$

$$S_w = \frac{1}{n} \sum_{i=1}^k S_i, \quad (3)$$

$$S_t = S_b + S_w. \quad (4)$$

The covariance matrix S_i of the i th class can be decomposed as $S_i = \tilde{A}_i \tilde{A}_i^T$, where each column of \tilde{A}_i corresponds to a data point from the i th class from which centroid m_i has been subtracted.

Define the matrices

$$H_w = \frac{1}{\sqrt{n}} [\tilde{A}_1, \dots, \tilde{A}_k], \quad (5)$$

$$H_b = [\sqrt{P_1}(m_1 - m), \dots, \sqrt{P_k}(m_k - m)].$$

Then, the scatter matrices S_w and S_b can be expressed as

$$S_b = H_b H_b^T, \quad S_w = H_w H_w^T. \quad (6)$$

The traces of the within-class and between-class scatter matrices can be computed as follows:

$$\text{trace}(S_b) = \sum_{i=1}^k P_i \|m_i - m\|^2,$$

$$\text{trace}(S_w) = \frac{1}{n} \sum_{i=1}^k \|\tilde{A}_i\|_F^2.$$

Hence, $\text{trace}(S_w)$ measures the closeness of the vectors within the classes, while $\text{trace}(S_b)$ measures the separation between the classes.

In the lower-dimensional space resulting from the linear transformation G , the within-class and between-class scatter matrices become

$$S_w^L = (G^T H_w)(G^T H_w)^T = G^T S_w G,$$

$$S_b^L = (G^T H_b)(G^T H_b)^T = G^T S_b G.$$

An optimal transformation G would maximize $\text{trace}(S_b^L)$ and minimize $\text{trace}(S_w^L)$. A common optimization for computing optimal G in classical LDA is

$$G = \arg \max_G \text{trace} \left((G^T S_w G)^{-1} G^T S_b G \right). \quad (7)$$

The solution can be readily computed by applying an eigen-decomposition on $S_w^{-1} S_b$, provided that the within-class scatter matrix S_w is nonsingular. Since the rank of the between-class scatter matrix is bounded from above by

$k - 1$, there are at most $k - 1$ discriminant vectors by classical LDA.

Note that a limitation of classical discriminant analysis in many applications involving small sample data, including gene expression data, is that the matrix S_w can be singular. This is known as the *undersampled* or *singularity* problems [20].

4 UNCORRELATED DISCRIMINANT ANALYSIS FOR NONSINGULAR S_w

Consider a variant of classical LDA in (7) as follows:

$$G = \arg \max_{G \in \mathbb{R}^{p \times t}: G^T S_t G = I_t} F(G), \quad (8)$$

where

$$F(G) = \text{trace} \left((G^T S_w G)^{-1} G^T S_b G \right).$$

From linear algebra, there exists a nonsingular matrix X such that

$$\begin{aligned} X^T S_w X &= I_p, \\ X^T S_b X &= \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p), \end{aligned} \quad (9)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

It can be shown that the matrix consisting of the first q columns of X (with normalizations) solves the optimization problem in (8), where q is the rank of the matrix S_b , as stated in the following theorem (the proof is attached in the Appendix).

Theorem 4.1. *Let the matrix X be defined in (9), and $q = \text{rank}(S_b)$. Let*

$$G^* = \left[\frac{1}{\sqrt{1 + \lambda_1}} X_1, \dots, \frac{1}{\sqrt{1 + \lambda_q}} X_q \right],$$

where X_i is the i th column of the matrix X and λ_i is defined in (9). Then, G^* solves the optimization problem in (8).

An efficient algorithm for computing $\{X_i\}_{i=1}^q$ through QR decomposition is given in Algorithm 1.

Algorithm 1. The computation of X

Input: the data matrix A

Output: the matrix X

1. Construct the matrices H_b and H_w as in (5).

2. Compute QR decomposition on H_w^T as

$$H_w^T = QR, \text{ where } Q \in \mathbb{R}^{n \times p}, R \in \mathbb{R}^{p \times p}.$$

3. Form the matrix $Y \leftarrow H_b^T R^{-1}$.

4. Compute SVD on Y as $Y = U \Sigma V^T$, where

$$U \in \mathbb{R}^{k \times q}, \Sigma \in \mathbb{R}^{q \times q}, V \in \mathbb{R}^{p \times q},$$

and $q = \text{rank}(H_b)$.

5. $X \leftarrow R^{-1} V$.

A nice property of the proposed LDA algorithm is that the features in the transformed space are uncorrelated, as stated in the following theorem (the proof is attached in the Appendix).

Theorem 4.2. *Let G^* be defined as in Theorem 4.1 and let Y be the original feature vector. The transformed feature of Y is denoted as $Z = G^T Y$. Let the i th feature component of Z is $Z_i = X_i^T Y$. Then, Z_i and Z_j are mutually uncorrelated, for any $i \neq j$.*

5 UNCORRELATED DISCRIMINANT ANALYSIS FOR THE GENERAL CASE

In Section 4, a variant of classical LDA is presented in (8). An efficient algorithm through QR decomposition is presented, provided that S_w is nonsingular. The key to the efficiency of the algorithm is the simultaneous diagonalization of the matrices S_w and S_b . In this section, we extend the algorithm to the case when S_w is singular.

In numerical linear algebra, a common way to diagonalize two matrices together is through the Generalized Singular Value Decomposition (GSVD). More details on GSVD can be found in [14].

Let

$$\Gamma = \begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix},$$

which is an $(n+k)$ by p matrix. By the Generalized Singular Value Decomposition, there exist orthogonal matrices $U \in \mathbb{R}^{k \times k}$, $V \in \mathbb{R}^{n \times n}$, and a nonsingular matrix $X \in \mathbb{R}^{p \times p}$, such that

$$\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^T \Gamma X = \begin{bmatrix} \Sigma_1 & 0 \\ \Sigma_2 & 0 \end{bmatrix}, \quad (10)$$

where

$$\Sigma_1 = \begin{bmatrix} I_b & 0 & 0 \\ 0 & D_b & 0 \\ 0 & 0 & 0_b \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0_w & 0 & 0 \\ 0 & D_w & 0 \\ 0 & 0 & I_w \end{bmatrix}.$$

Here,

$$I_b \in \mathbb{R}^{r \times r} \text{ and } I_w \in \mathbb{R}^{(t-r-s) \times (t-r-s)}$$

are identity matrices,

$$0_b \in \mathbb{R}^{(k-r-s) \times (t-r-s)} \text{ and } 0_w \in \mathbb{R}^{(n-t+r) \times r}$$

are zero matrices with

$$\begin{aligned} r &= \text{rank}(\Gamma) - \text{rank}(H_w), \\ s &= \text{rank}(H_b) + \text{rank}(H_w) - \text{rank}(\Gamma), \end{aligned}$$

and

$$\begin{aligned} D_b &= \text{diag}(\alpha_{r+1}, \dots, \alpha_{r+s}), \\ D_w &= \text{diag}(\beta_{r+1}, \dots, \beta_{r+s}) \end{aligned}$$

are diagonal matrices satisfying

$$\begin{aligned} 1 &> \alpha_{r+1} \geq \dots \geq \alpha_{r+s} > 0, \\ 0 &< \beta_{r+1} \leq \dots \leq \beta_{r+s} < 1, \end{aligned}$$

and $\alpha_i^2 + \beta_i^2 = 1$ for $i = r+1, \dots, r+s$.

From (10), we have

$$\begin{aligned} H_b^T X &= U[\Sigma_1 \quad 0] \\ H_w^T X &= V[\Sigma_2 \quad 0] \end{aligned}$$

and

$$\begin{aligned} X^T H_b H_b^T X &= \begin{bmatrix} \Sigma_1^T \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \equiv D_1, \\ X^T H_w H_w^T X &= \begin{bmatrix} \Sigma_2^T \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} \equiv D_2. \end{aligned}$$

Hence, a natural extension of the proposed LDA algorithm in Section 4 is to choose the first $q = r+s$ columns of the matrix X in (10) as the transformation matrix G . Note that no normalization is required here since $\Sigma_1^T \Sigma_1 + \Sigma_2^T \Sigma_2$ is an identity matrix. The main algorithm is given in Algorithm 2.

Algorithm 2. The ULDA Algorithm

Input: the data matrix A

Output: the transformation matrix G

1. Form the matrices H_b and H_w as in Eq. (5).
2. Compute GSVD on the matrix pair (H_b^T, H_w^T) to obtain the matrix X as in Eq. (10).
3. $q \leftarrow \text{rank}(H_b)$.
4. $G \leftarrow [X_1, \dots, X_q]$.

The following result is straightforward from Theorem 4.1.

Theorem 5.1. *Let $G^* = [X_1, \dots, X_q]$, where X_i is computed by GSVD as in (10), and let Y be the original feature vector. The transformed feature of Y is denoted as $Z = G^T Y$. Let the i th feature component of Z is $Z_i = X_i^T Y$. Then, Z_i and Z_j are mutually uncorrelated, for any $i \neq j$.*

6 EXPERIMENTAL RESULTS AND ANALYSIS

The ULDA algorithm is implemented in MATLAB. The source code and all the data sets used in the following experiments can be accessed at <http://www.cs.umn.edu/~jieping/ULDA>.

6.1 The Data Sets

We use a wide range of publicly available data sets and we expect these data sets would provide us enough insights on the behavior of ULDA.

- The ALL/AML database contains the gene expression profiles of two acute cases of leukemia: acute lymphoblastic leukemia (ALL for short) and acute myeloblastic leukemia (AML for short). The ALL part of the data set comes from two sample types, B-cell and T-cell, and the AML part is split into bone marrow samples and peripheral blood. This data set was first studied in the seminal paper of Golub et al. [15]. Golub et al. studied this problem to address the binary

classification problem between the AML samples and the ALL samples. However, due to the bipartition of each component, it can be treated as a three-class data set (*B-cell*, *T-cell*, and *AML*) or as a four-class data set (*B-cell*, *T-cell*, *AML-BM*, and *AML PB*). The three-class version is referred to here as **ALL-AML-3** and the four-class version as **ALL-AML-4**.

- The **LEUKEMIA** data set comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML), and is available at <http://www.genome.wi.mit.edu/MPR>. It was studied in [8], [15].
- The **ALL** data set [34] covers six subtypes of acute lymphoblastic leukemia. The data set is available at <http://www.stjude.com/research/data/ALL1>.
- The **GCM** data set consists of 198 human tumor samples spanning fourteen different cancer types. The data set was first studied in [24], [33].
- **SRBCT** [19] is the data set of small, round blue cell tumors of childhood and can be downloaded at <http://research.nhgri.nih.gov/microarray/Supplement>. This data set consists of 83 samples spanning four classes (excluding five samples as done in [27]).
- **LYMPHOMA** is a data set of the three most prevalent adult lymphoid malignancies and is available at <http://genome-www.stanford.edu/lymphoma>. The data set was first studied in [1].
- The **COLON** data set in [2] consists of 40 tumor and 22 normal colon tissues. The data set is available at <http://microarray.princeton.edu/oncology>.
- The **PROSTATE** data set in [26] contains 52 prostate tumor samples and 50 nontumor prostate samples. It is available at <http://www.broad.mit.edu/cgi-bin/cancer/data/sets.cgi>.

The characteristics of the above data sets are summarized in Table 2.

We performed our benchmark study by repeated random splitting into learning and test sets exactly as in [8]. The data was partitioned randomly into a learning set consisting of two-thirds of the whole set and a test set consisting of one-third of the whole set. To reduce the variability, the splitting was repeated 50 times and the resulting accuracies were averaged.

6.2 Experimental Setup

We used the classification accuracy on the test set as the performance measure. All of our experiments were performed on a P4 2GHz machine with 512M memory running Linux 2.4.9-31. Classification is done via the K-Nearest Neighbor (KNN) algorithm with $K = 1$.

6.3 Gene Selection

As we discussed in Section 3, ULDA sets an upper limit on the number dimensions in the transformed space due to the rank issues. For KNN and DLDA, the classification results from using selected features are reported. The feature selection is based on the *BW* ratio described in [8]. Basically, for a gene j , the *BW* ratio is

$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)(m_k(j) - m(j))^2}{\sum_i \sum_k I(y_i = k)(a_{ij} - m_k(j))^2},$$

TABLE 2
Summary of the Data Sets Used in the Experiments

Dataset	sample size	# genes	# classes
ALL-AML-3	72	7129	3
ALL-AML-4	72	7129	4
LEUKEMIA	72	7129	2
ALL	248	12558	6
GCM	198	16063	14
SRBCT	83	2308	4
LYMPHOMA	62	4026	3
COLON	62	2000	2
PROSTATE	102	6033	2

where $m(j)$ and $m_k(j)$ denote the average expression level of gene j across all samples and across samples belonging to class k only, y_i is the class label for the i th sample, and a_{ij} is the expression level of gene j in the i th sample. $I(\cdot)$ denotes the indicator function, which equals 1 if the condition of the indicator function in the parenthesis is true and 0 otherwise. The number of genes selected is based on cross-validation. All the genes in the data set are used in ULDA and SVM and, thus, gene selection was not applied for these two methods.

6.4 Experimental Results

Gene selection is applied as the preprocessing step for both KNN and DLDA, where the optimal number of selected genes is determined by cross-validation. Fig. 1 shows the effect of gene selection on the classification performance of KNN (left) and DLDA (right). The horizontal axis denotes the data sets and the vertical axis denotes the classification accuracy. We can observe that gene selection improves the performance of both KNN and DLDA. The improvement is more significant for DLDA, which is related to the estimation scheme of the covariance matrix applied in DLDA. We also applied gene selection for ULDA and SVM, and the results show that the improvement of classification accuracy is small. In the following experiments, we report the results on KNN and DLDA with gene selection.

The number of reduced dimensions in ULDA is set to be $q = \text{rank}(H_b)$. (See line 4 of Algorithm 2.) However, the last few dimensions may have been contributed by the noise in the data set. To verify this, we run ULDA by keeping the first \tilde{q} dimensions only, where $1 \leq \tilde{q} \leq q$. The classification results on the GCM data set are shown in Fig. 2, where the horizontal axis is the number of reduced dimensions and the vertical axis is the classification accuracy. We can observe that the accuracy monotonically increases when the number of reduced dimensions increases, until the maximum possible value $q = \text{rank}(H_b)$ is reached. Similar trends have been observed from other data sets, so the results are omitted. In the following experiment, we keep all the q dimensions for ULDA.

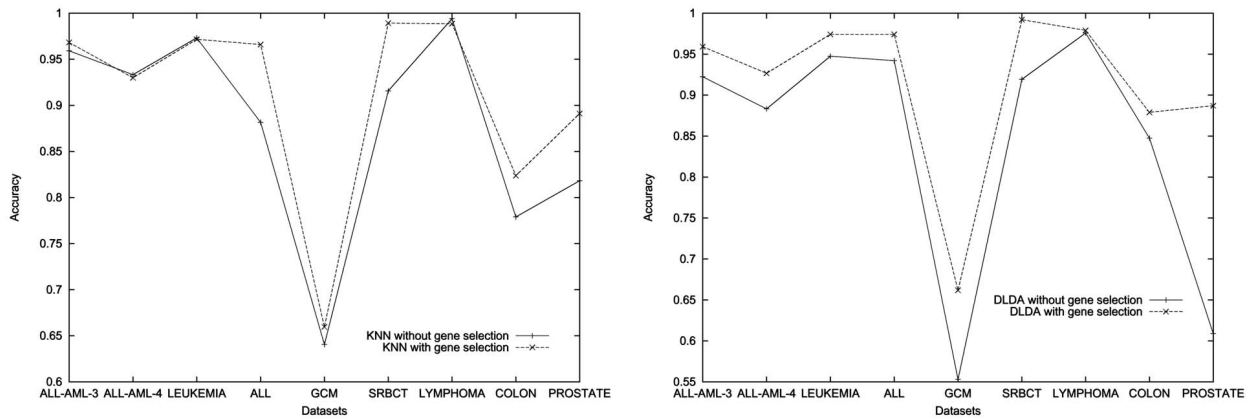


Fig. 1. Effect of gene selection on the classification performance of KNN and DLDA.

Table 3 presents the experimental results and show performance comparisons. Boxplots of classification accuracies for different algorithms are presented in Figs. 3, 4, 5, 6, and 7. It can be seen that ULDA is close to the best classifier in all cases. We can also observe that for many data sets, the performance of different algorithms are close to one another. However, the optimal values of the parameters involved in KNN, DLDA, NSC, and SVM seem to be problem-dependent. It may be expensive to choose the optimal parameters via cross-validation, whereas ULDA has the advantage of not involving any parameters.

We showed theoretically in Section 4 and Section 5 that the features extracted by ULDA are uncorrelated, i.e., ULDA removes the redundancy in the original features while keeping the most discriminative information. This theoretical result is further confirmed by the comparative results in Table 3. For most data sets used in the experiments, the original dimensions are in the thousands or tens of thousands, while the reduced dimensions by ULDA are less than 10 in most cases. ULDA is able to extract a small number of discriminative features without loss of classification accuracy (ULDA is competitive to KNN in all data sets).

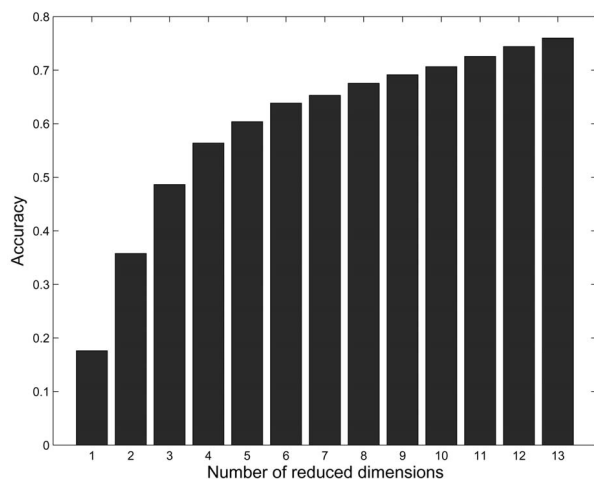


Fig. 2. Effect of the number of reduced dimensions on the classification performance of ULDA for the GCM data set.

Let us take a closer look at DLDA. The key difference between DLDA and ULDA is that DLDA only uses a subset of genes, while ULDA utilizes the information from all the genes and removes the redundant features. As mentioned in Section 1, in expression data, many gene groups interact closely. Gene interactions are important biologically and may contribute to class distinctions. This is confirmed by the competitive performance of ULDA over DLDA. However, it is interesting to note that a simple classifier such as DLDA performs reasonably well. For large and high-dimensional gene expression data, DLDA is advantageous, due to its high efficiency.

The competitive performance of ULDA over SVM is partly due to the high dimension of the gene expression data. In general, high dimension leads to high possibility of linear separability between different classes. It is worthwhile to note that SVM involves the tuning of certain parameters, which are usually data-dependent and can be expensive, while ULDA is parameter-free.

We now take a close look at the results obtained by ULDA. On ALL-AML-3 and ALL-AML-4, errors were made in distinguishing between the subtypes of ALL (*B-cell* and *T-cell*) and the subtypes of AML (*AML-BM* and *AML-PB*). This is understandable due to the extremely small size of the subtypes. The performance on GCM is not good for all the classifiers. As discussed in [24], multiclass distinctions among highly related tumor types are intrinsically difficult. The low accuracy is probably due to the small sample size versus the relative large number of classes. In addition, the effects of biological and measurement noise, contaminating nonmalignant tumor components, and inclusion of genetically heterogeneous samples within clinically defined tumor classes may all effectively decrease predictive power [24]. Increasing both gene and sample numbers would be a plausible solution for better prediction.

In summary, our ULDA algorithm is an effective method for tissue classification. The experimental results along with the theoretical proofs show that ULDA provides a good choice for practical classification problems.

TABLE 3
Comparison of Classification Accuracy Rate for Five Classifiers and Nine Data Sets

Datasets/Methods	ULDA	KNN	DLDA	NSC	SVM
ALL-AML-3	96.75(2.83)	96.83(3.53)	95.92(3.52)	95.25(3.67)	96.25(4.06)
ALL-AML-4	91.42(5.73)	93.00(5.92)	92.67(5.62)	93.50(4.47)	92.58(4.32)
LEUKEMIA	97.67(2.40)	97.17(3.31)	97.42(2.90)	95.25(3.67)	97.50(2.23)
ALL	97.20(2.01)	96.60(1.83)	97.42(1.44)	96.65(2.06)	97.23(1.64)
GCM	75.05(6.32)	65.97(4.90)	66.19(5.25)	70.06(6.30)	70.31(5.15)
SRBCT	100.0(0.00)	98.93(2.42)	99.21(2.20)	97.71(2.37)	100.0(0.00)
LYMPHOMA	99.24(1.76)	98.86(2.46)	97.91(3.07)	96.57(5.61)	99.86(0.67)
COLON	85.24(5.46)	82.38(6.68)	87.90(5.29)	86.29(6.64)	85.05(6.94)
PROSTATE	92.04(3.67)	89.12(4.54)	88.71(5.68)	90.53(4.58)	92.53(4.04)

The mean and standard deviation (in parenthesis) of accuracies from 50 runs are reported.

7 CONCLUSIONS

In this paper, we have presented an algorithm, called ULDA for gene expression data classification. The features in the transformed spaces of ULDA are uncorrelated. In addition, ULDA utilizes GSVD to handle the undersampled problems. As a result, ULDA shows good discriminating power in gene expression data analysis. Extensive experiments clearly demonstrate its effectiveness.

There are several natural avenues for future research. First, we could incorporate prediction strength [15] in ULDA. The prediction strength can be naturally defined as a function of the distance between the new instance and the class center in the transformed space of ULDA. Second, since the features in the transformed space obtained via ULDA are uncorrelated, combining other methods with ULDA is a promising direction for further investigation.

Third, ULDA basically performs dimension reduction by transforming the original data into a low-dimensional space. In the transformed space, each feature is a linear combination of the original features (genes). We plan to explore the relationship between the coefficients of the transformation matrix and the discriminating power of genes.

APPENDIX

Proof of Theorem 4.1. It is clear that $(G^*)^T S_i G^*$ is an identity matrix, i.e., the constraint in (8) is satisfied. Next, we only need to show that the minimum of $F(G)$ is obtained at G^* .

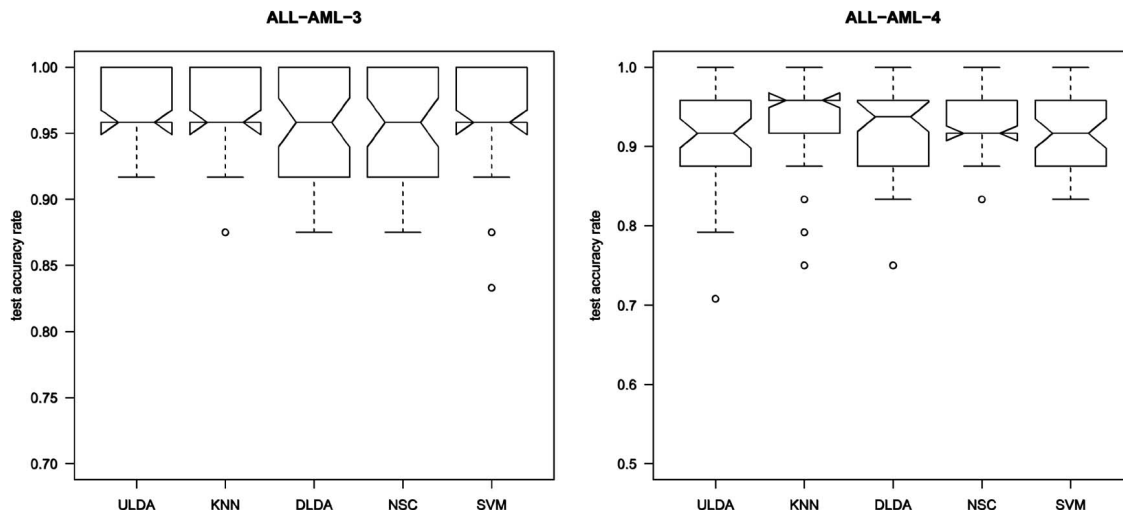


Fig. 3. Boxplots of classification accuracies for Uncorrelated Linear Discriminant Analysis (ULDA), K-Nearest Neighbor (KNN), Diagonal Linear Discriminant Analysis (DLDA), Nearest Shrunken Centroid (NSC), and Support Vector Machines (SVM) on the ALL-AML-3 and ALL-AML-4 data sets.

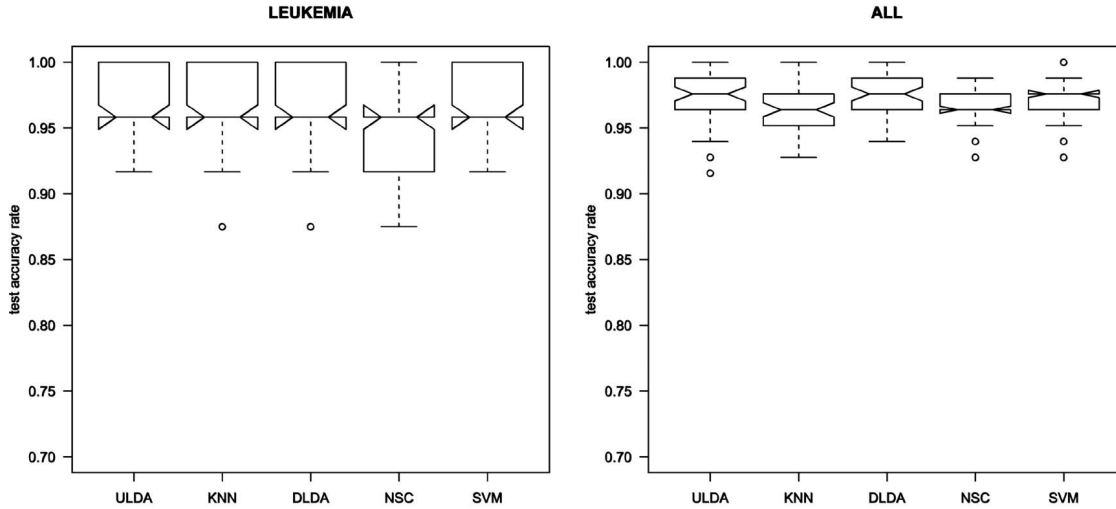


Fig. 4. Boxplots of classification accuracies for ULDA, KNN, DLDA, NSC, and SVM on the **LEUKEMIA** and **ALL** data sets.

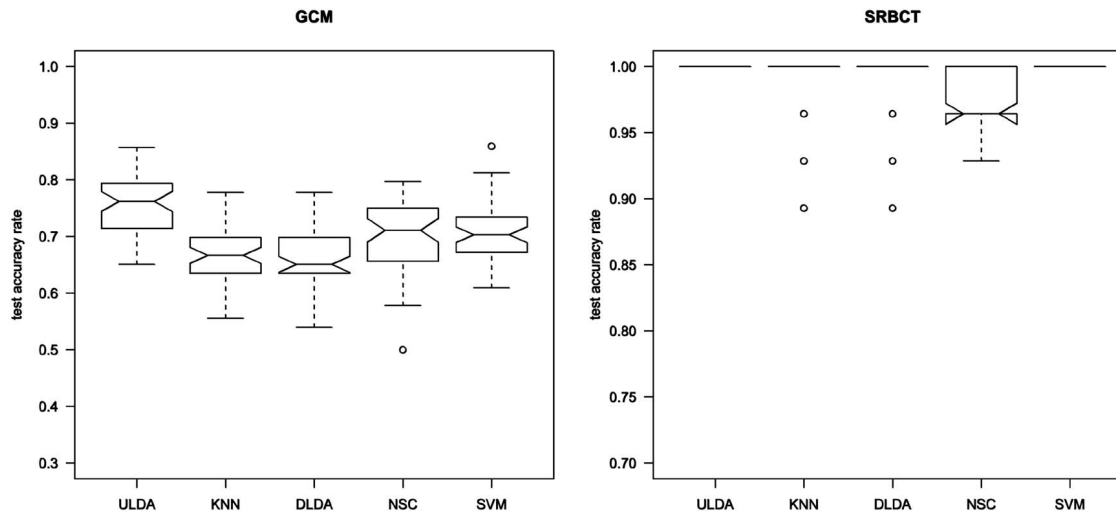


Fig. 5. Boxplots of classification accuracies for ULDA, KNN, DLDA, NSC, and SVM on the **GCM** and **SRBCT** data sets.

$$\begin{aligned} F(G) &= \text{trace}\left(\left(G^T S_w G\right)^{-1} \left(G^T S_b G\right)\right) \\ &= \text{trace}\left(\left(\tilde{G} \tilde{G}^T\right)^{-1} \left(\tilde{G} \Lambda \tilde{G}^T\right)\right), \end{aligned}$$

where $\tilde{G} = (X^{-1}G)^T$ and Λ is defined in (9). Let $\tilde{G}^T = QR$ be the reduced QR factorization of \tilde{G}^T , where $Q \in \mathbb{R}^{p \times \ell}$ has orthonormal columns and R is nonsingular. Using the fact that $\text{trace}(AB) = \text{trace}(BA)$, for any matrices A and B , we have

$$\begin{aligned} F(G) &= \text{trace}\left(\left(R^T R\right)^{-1} \left(R^T Q^T \Lambda Q R\right)\right) \\ &= \text{trace}\left(R^{-1} Q^T \Lambda Q R\right) \\ &= \text{trace}\left(Q^T \Lambda Q R R^{-1}\right) \\ &= \text{trace}\left(Q^T \Lambda Q\right) \leq \lambda_1 + \dots + \lambda_q, \end{aligned}$$

where the inequality becomes equality for

$$Q = \begin{pmatrix} I_\ell \\ 0 \end{pmatrix} \text{ or } G = X \begin{pmatrix} I_\ell \\ 0 \end{pmatrix} R,$$

when the reduced dimension $\ell = q$. Note that R is an arbitrary nonsingular matrix and G^* corresponds to the case when we choose R to be

$$R = \text{diag}\left(\frac{1}{\sqrt{1+\lambda_1}}, \dots, \frac{1}{\sqrt{1+\lambda_q}}\right).$$

This completes the proof of the theorem. \square

Proof of Theorem 4.2. It is easy to check that the covariance between Z_i and Z_j can be computed as

$$\begin{aligned} \text{Cov}(Z_i, Z_j) &= E(Z_i - EZ_i)(Z_j - EZ_j) \\ &= X_i^T \{E(Y - EY)(Y - EY)^T\} X_j \\ &= X_i^T S_t X_j. \end{aligned} \quad (11)$$

Hence, their correlation coefficient is

$$\text{Cor}(Z_i, Z_j) = \frac{X_i^T S_t X_j}{\sqrt{X_i^T S_t X_i} \sqrt{X_j^T S_t X_j}}. \quad (12)$$

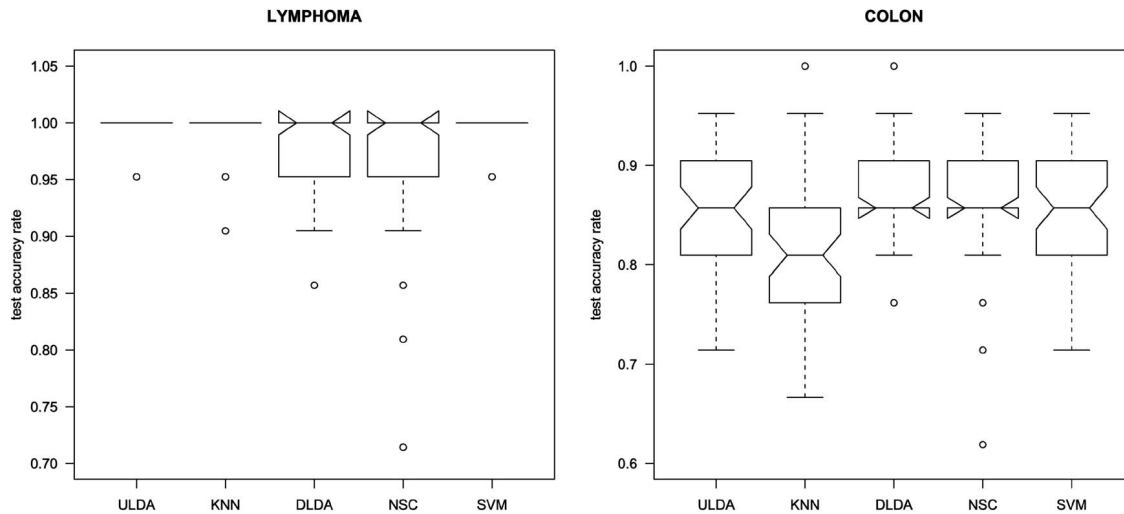


Fig. 6. Boxplots of classification accuracies for ULDA, KNN, DLDA, NSC, and SVM on the **LYMPHOMA** and **COLON** data sets.

Since $X_i^T S_i X_j = 0$, for $i \neq j$, we have $\text{Cor}(Z_i, Z_j) = 0$, for $i \neq j$. That is, the components of the vectors transformed by the proposed LDA algorithm are mutually uncorrelated. This completes the proof of the theorem. \square

ACKNOWLEDGMENTS

The authors would like to thank the associate editor and the reviewers for helpful comments that greatly improved the paper. This research is sponsored, in part, by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. The research of J. Ye is also supported by fellowships from Guidant Corporation and from the Department of Computer Science and Engineering at the University of Minnesota.

REFERENCES

- [1] A.A. Alizadeh, M.B. Eisen, R.E. David, C. Ma, I.S. Lossos, A. Rosenwald, H.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Martu, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, G.P. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botsten, P.O. Brown, and L.M. Staudt, "Distinct Types Of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, vol. 403, pp. 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 Arrays," *Proc. Nat'l Academy of Science*, vol. 96, pp. 6745-6750, 1999.
- [3] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue Classification with Gene Expression Profiles," *J. Computational Biology*, vol. 7, pp. 559-584, 2000.
- [4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997.
- [5] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, and T.S. Furey, "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proc. Nat'l Academy of Science*, vol. 97, pp. 262-267, 2000.
- [6] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [7] M. Chee, R. Yang, E. Hubbell, A. Berno, X. Huang, D. Stern, J. Winkler, D. Lockhart, M. Morris, and S. Fodor, "Accessing Genetic Information with High Density DNA Arrays," *Science*, vol. 274, pp. 610-614, 1996.
- [8] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.*, vol. 97, pp. 77-87, 2002.
- [9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Science*, vol. 95, pp. 14863-4868, 1998.
- [10] S. Fodor, J. Read, M. Pirrung, L. Stryer, A. Lu, and D. Solas, "Light-Directed, Spatially Addressable Parallel Chemical Synthesis," *Science*, vol. 251, pp. 767-783, 1991.
- [11] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *J. Computational Biology*, vol. 7, pp. 601-620, 2000.
- [12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
- [13] G. Getz, E. Levine, and E. Domany, "Coupled Two-Way Clustering Analysis of Gene Microarray Data," *Proc. Nat'l Academy of Science*, vol. 97, pp. 12079-12084, 2000.

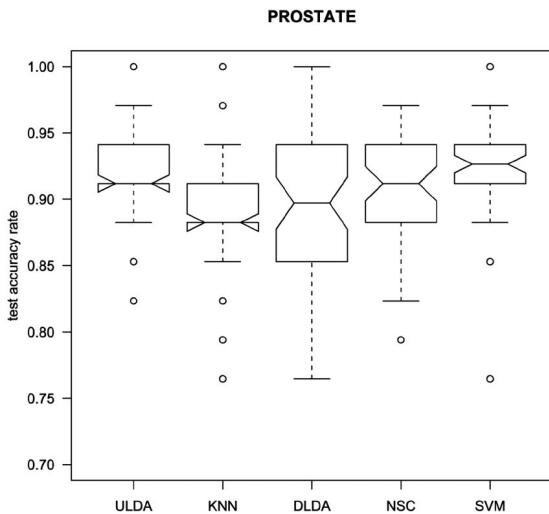


Fig. 7. Boxplots of classification accuracies for ULDA, KNN, DLDA, NSC, and SVM on the **PROSTATE** data set.

- [14] G.H. Golub and C.F. V. Loan, *Matrix Computations*. The Johns Hopkins Univ. Press, 1991.
- [15] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, 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*, vol. 286, pp. 531-537, 1999.
- [16] T. Hastie, R. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [17] P. Howland, M. Jeon, and H. Park, "Structure Preserving Dimension Reduction for Clustered Text Data Based on the Generalized Singular Value Decomposition," *SIAM J. Matrix Analysis and Applications*, vol. 25, no. 1, pp. 165-179, 2003.
- [18] C.W. Hsu and C.J. Lin, "A Comparison of Methods for Multi-Class Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415-425, 2002.
- [19] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, pp. 673-679, 2001.
- [20] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas, "Discriminant Analysis with Singular Covariance Matrices: Methods and Applications to Spectroscopic Data," *Applied Statistics*, vol. 44, pp. 101-115, 1995.
- [21] Y. Lee and C.K. Lee, "Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data," *Bioinformatics*, vol. 19, no. 9, pp. 1132-1139, 2003.
- [22] 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*, vol. 20, no. 15, pp. 2429-2437, 2004.
- [23] C. Ooi and P. Tan, "Genetic Algorithms Applied to Multi-Class Prediction for the Analysis of Gene Expression Data," *Bioinformatics*, vol. 19, pp. 37-44, 2003.
- [24] 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 R.T. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures," *Proc. Nat'l Academy of Science*, vol. 98, pp. 15149-15154, 2001.
- [25] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellmand, V. Iyer, S.S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, and M.P.O. Brown, "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, vol. 24, pp. 227-235, 2000.
- [26] D. Singh et al., "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203-209, 2002.
- [27] 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*, 2004.
- [28] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Science*, vol. 99, no. 10, pp. 6567-6572, 2002.
- [29] V.N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
- [30] J. Ye, "Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems," pending publication.
- [31] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature Extraction via Generalized Uncorrelated Linear Discriminant Analysis," *Proc. 21st Int'l Conf. Machine Learning*, pp. 895-902, 2004.
- [32] J. Ye, R. Janardan, C.H. Park, and H. Park, "An Optimization Criterion for Generalized Discriminant Analysis on Undersampled Problems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982-994, Aug. 2004.
- [33] C.H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. Rifkin, M. Angelo, M. Reich, E.S. Lander, J.P. Mesirov, and T.R. Golub, "Molecular Classification of Multiple Tumor Types," *Bioinformatics*, vol. 11, pp. 1-7, 2001.

- [34] E.-J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahrouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W.E. Evans, C. Naeve, L. Wong, and J.R. Downing, "Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Lymphoblastic Leukemia by Gene Expression Profiling," *Cancer Cell*, vol. 1, pp. 133-143, 2002.
- [35] K.Y. Yeung and W.L. Ruzzo, "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, vol. 17, pp. 763-774, 2001.



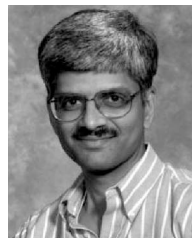
Jieping Ye received the BS degree in mathematics from Fudan University, Shanghai, China, in 1997. He is currently a PhD student in the Department of Computer Science and Engineering, University of Minnesota. He was awarded the Guidant Fellowship in 2004-2005. In 2004, his paper on generalized low rank approximations of matrices won the outstanding student paper award at the 21st International Conference on Machine Learning. His research interests include data mining, machine learning, pattern recognition, bioinformatics, and geometric modeling.



Tao Li received the PhD degree in computer science from the University of Rochester in 2004. He is currently an assistant professor in the School of Computer Science at Florida International University. His primary research interests are: data mining, machine learning, bioinformatics, and music information retrieval.



Tao Xiong received the MEng degree in electronics and information engineering from Beijing Jiaotong University, Beijing, China, in 2000. He is currently a PhD student in the Department of Electrical and Computer Engineering at the University of Minnesota, Twin Cities. His research interests include machine learning, data mining, bioinformatics, biomedical imaging, and signal processing.



Ravi Janardan received the PhD degree in computer science from Purdue University in 1987. He is a professor and associate head in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. His research interests are in the design and analysis of geometric algorithms and data structures, and their application to problems in a variety of areas, including computer-aided design and manufacturing, transportation, VLSI design, bioinformatics, and computer graphics. He has published extensively in these areas.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.