Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection

Satoshi Niijima and Yasushi Okuno

Abstract—Until recently, numerous feature selection techniques have been proposed and found wide applications in genomics and proteomics. For instance, feature/gene selection has proven to be useful for biomarker discovery from microarray and mass spectrometry data. While supervised feature selection has been explored extensively, there are only a few unsupervised methods that can be applied to exploratory data analysis. In this paper, we address the problem of unsupervised feature selection. First, we extend Laplacian linear discriminant analysis (LLDA) to unsupervised cases. Second, we propose a novel algorithm for computing LLDA, which is efficient in the case of high dimensionality and small sample size as in microarray data. Finally, an unsupervised feature selection method, called LLDA-based Recursive Feature Elimination (LLDA-RFE), is proposed. We apply LLDA-RFE to several public data sets of cancer microarrays and compare its performance with those of Laplacian score and SVD-entropy, two state-of-the-art unsupervised methods, and with that of Fisher score, a supervised filter method. Our results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. It performs even better than Fisher score for some of the data sets, despite the fact that LLDA-RFE is fully unsupervised.

Index Terms—Unsupervised feature selection, linear discriminant analysis, graph Laplacian, microarray data analysis.

1 Introduction

In recent years, feature/gene selection methods have been widely used in genomics and proteomics to handle a deluge of data produced by high-throughput technologies such as microarray and mass spectrometry. In microarray studies, for instance, a small fraction of genes typically exhibit significant differential expression among tens of thousands of genes whose expression levels are measured simultaneously. Thus, it is of great importance to identify genes relevant to a biological phenomenon of interest and to characterize their expression profiles. Gene selection can be useful for multiple purposes: to save computational costs of subsequent analysis by reducing the number of genes, to improve the prediction performance of classifiers by using discriminative genes only, and to identify informative genes for further investigation of their biological relevance. Specifically, gene selection has proven to be useful for biomarker discovery in cancer studies, i.e., searching for potential marker genes contributing to classification of cancer subtypes or prediction of clinical outcomes, which leads to more reliable diagnosis and better treatments of cancer.

To date, numerous techniques for feature selection have been developed [12] and also applied successfully to the analysis of biological data with many features. In contrast to supervised feature selection, however, unsupervised feature selection has not yet been explored extensively. Indeed, there have been only a few unsupervised methods proposed until recently [7], [14], [15], [28], [30]. Unsupervised feature selection is of great use in particular for class discovery. For instance, clustering is usually performed to find clusters in microarray samples on the basis of the expression profiles of all genes, but the clusters so obtained can be obscured by the large number of irrelevant genes. Therefore, unsupervised feature selection is essential to the exploratory analysis of biological data. Moreover, even when class labels are provided by external knowledge, but may be unreliable or mislabeled, overfitting can be alleviated by performing feature selection in an unsupervised manner. It is obviously more challenging to identify features that reveal underlying cluster structures in the samples than to find those exhibiting similar patterns across all the samples.

To address this problem, we propose an unsupervised feature selection method, called Laplacian linear discriminant analysis-based recursive feature elimination (LLDA-RFE). LLDA-RFE is closely related to Laplacian score [15], which is also based on graph Laplacian and can be applied in an unsupervised manner. The major difference is that, whereas Laplacian score is a univariate approach, LLDA-RFE is multivariate, allowing for selecting features that contribute to discrimination in combination with other features. Recently, Wolf and Shashua [30] proposed the $Q-\alpha$ algorithm, which takes advantage of the spectral properties of the graph Laplacian of features. While the $Q-\alpha$ algorithm has an interesting property that the sparsity of features naturally emerges, it does not scale well to the feature size. Also, the algorithm involves iterative computations on a matrix of the feature size in a least-squares optimization process to ensure a local maximum solution. In contrast, our proposed algorithm for LLDA-RFE is computationally tractable and has a global maximum solution. It is shown that Laplacian linear

• *The authors are with the Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan.*
E-mail: {niijima, okuno}@pharm.kyoto-u.ac.jp.

Manuscript received 2 May 2007; revised 21 Aug. 2007; accepted 28 Sept. 2007; published online 11 Oct. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-05-0052. Digital Object Identifier no. 10.1109/TCBB.2007.70257.

discriminant analysis (LLDA) includes the maximum margin criterion (MMC) [18] as a supervised case. Although LLDA-RFE is a natural extension of MMC-RFE, the proposed algorithm needs not reduce dimensionality before applying LLDA, unlike the MMC-RFE algorithm proposed previously [24].

We compare the performance of LLDA-RFE with those of Laplacian score and SVD-entropy [28], two state-of-the-art unsupervised feature selection methods, on seven public data sets of cancer microarrays. The performances of these methods are evaluated by their capability of identifying discriminative genes without using class information. We also compare the performance between LLDA-RFE and Fisher score [8], [15], a supervised filter method. Experimental results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. Despite the fact that LLDA-RFE is fully unsupervised, it performs even better than Fisher score for some of the data sets.

The rest of this paper is organized as follows: In Section 2, we give outlines of linear discriminant analysis (LDA) and the MMC. We then introduce LLDA and extend it to unsupervised cases in Section 3. An efficient algorithm for computing LLDA is also proposed. We present the LLDA-RFE algorithm for feature selection in Section 4. Section 5 describes related work on unsupervised feature selection. Experimental results on seven microarray data sets are presented and discussed in Section 6. Finally, we give concluding remarks in Section 7.

2 LDA AND MMC

In this section, we outline LDA and the MMC as preliminaries to the introduction of supervised LLDA, which is then extended to unsupervised LLDA in Section 3.

LDA aims to find a set of projection vectors that maximize the between-class scatter and simultaneously minimize the within-class scatter, thereby achieving maximum discrimination [9].

Let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ be a sample matrix containing $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ as columns, where p is the number of features, and n is the number of samples. The between-class scatter matrix $\boldsymbol{S}_b$ and the within-class scatter matrix $\boldsymbol{S}_w$ are defined as

$$\boldsymbol{S}_b = \frac{1}{n}\sum_{k=1}^{c} n_k\Big(\boldsymbol{m}^{(k)} - \boldsymbol{m}\Big)\Big(\boldsymbol{m}^{(k)} - \boldsymbol{m}\Big)^T,$$

$$\boldsymbol{S}_w = \frac{1}{n}\sum_{k=1}^{c}\sum_{j=1}^{n_k}\Big(\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)}\Big)\Big(\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)}\Big)^T,$$

where c is the number of classes, n_k is the number of samples in class k, $\boldsymbol{x}_j^{(k)}$ is the jth sample in class k, and $\boldsymbol{m}^{(k)}$ and $\boldsymbol{m}$ are the mean vector of class k and the total mean vector, respectively. Then, classical LDA finds the projection matrix $\boldsymbol{W}$ that maximizes the Fisher criterion

$$J_{LDA}(\boldsymbol{W}) = trace\Big((\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W})^{-1}(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W})\Big), \tag{1}$$

subject to, e.g., the orthogonality constraint on $\boldsymbol{W}$, i.e., $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$. By solving a generalized eigenvalue problem, $\boldsymbol{W}$ can be found as the eigenvectors of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ corresponding to the largest eigenvalues. However, when the dimensionality of samples is larger than the sample size, i.e., $p > n$, $\boldsymbol{S}_w$ becomes singular and we cannot compute $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$, which is a major drawback of classical LDA. This is known as the singularity problem or the small sample size problem.

To overcome this problem, Li et al. [18] proposed to use the MMC instead of (1) to find the projection vectors. The MMC is defined as

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \boldsymbol{S}_w)\boldsymbol{W}\big). \tag{2}$$

In this case, the projection matrix $\boldsymbol{W}$ that maximizes (2) can be found as the eigenvectors of $\boldsymbol{S}_b - \boldsymbol{S}_w$ corresponding to the largest eigenvalues. Li et al. proposed an efficient algorithm to compute the projection matrix of the MMC under the constraint that $\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W} = \boldsymbol{I}$, where $\boldsymbol{S}_t$ is the total scatter matrix defined as

$$\boldsymbol{S}_t = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T.$$

Li's algorithm is found to be the same as the uncorrelated LDA (ULDA) algorithm in [32]. Also, an efficient algorithm for the MMC subject to the orthogonality constraint that $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$ was presented in [24]. In both cases, we need not compute the inverse of $\boldsymbol{S}_w$; hence, the singularity problem can be easily avoided.

It should be noted that the MMC is not equivalent to the Fisher criterion. The discriminant vectors obtained by maximizing (2) are not generally the same as those obtained by maximizing (1) [21]. More precisely, although ULDA can be considered as an extension of classical LDA to small sample size cases [32], the MMC with the orthogonality constraint does not necessarily yield projection vectors that are optimal for discrimination. In practice, a better discrimination can be achieved by balancing the between-class and within-class scatters using the following criterion as in [20]:

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \mu\boldsymbol{S}_w)\boldsymbol{W}\big), \tag{3}$$

where μ is a nonnegative constant. It is clear that (2) is a special case of (3). For simplicity, the present study focuses on the MMC defined by (2) with the orthogonality constraint.

3 UNSUPERVISED LLDA

3.1 Extension of LLDA to Unsupervised Cases

We can rewrite the total and within-class scatter matrices as follows:

$$\begin{aligned}\boldsymbol{S}_t &= \frac{1}{n}\boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T\right)\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_g)\boldsymbol{X}^T, \\ \boldsymbol{S}_w &= \frac{1}{n}\boldsymbol{X}\left(\boldsymbol{I} - \sum_{k=1}^{c}\frac{1}{n_k}\boldsymbol{e}^{(k)}\boldsymbol{e}^{(k)^T}\right)\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_\ell)\boldsymbol{X}^T,\end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{e} = (1, 1, \ldots, 1)^T$ is an n-dimensional vector, and $\boldsymbol{e}^{(k)}$ is an n-dimensional vector

with $e_i^{(k)} = 1$ if $\boldsymbol{x}_i$ belongs to class k, and 0 otherwise. In terms of graph Laplacians [6], $\boldsymbol{I} - \boldsymbol{W}_g$ can be viewed as the Laplacian of a global graph such that all vertices are connected each other with a constant weight of $1/n$, and $\boldsymbol{I} - \boldsymbol{W}_\ell$ as the Laplacian of a local graph such that a pair of vertices are connected with a constant weight of $1/n_k$ only when both belong to the kth class. In the following, we refer to the Laplacian of a globally connected graph as the global Laplacian and the Laplacian of a locally connected graph as the local Laplacian.

From relationship [9]

$$\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_w,$$

it follows that

$$\begin{aligned} \boldsymbol{S}_b - \boldsymbol{S}_w &= \boldsymbol{S}_t - 2\boldsymbol{S}_w \\ &= \frac{1}{n}\boldsymbol{X}\big((\boldsymbol{I} - \boldsymbol{W}_g) - 2(\boldsymbol{I} - \boldsymbol{W}_\ell)\big)\boldsymbol{X}^T. \end{aligned} \tag{4}$$

The MMC represented in this form is referred to as LLDA in [26] and was applied to extract discriminant features in supervised scenarios.

In this study, we extend (4) to unsupervised cases. We first define the global similarity matrix $\boldsymbol{K}_g$ and the local similarity matrix $\boldsymbol{K}_\ell$ as

$$[\boldsymbol{K}_g]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } i \neq j, \\ 0, & \text{otherwise,} \end{cases}$$

$$[\boldsymbol{K}_\ell]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } \boldsymbol{x}_i \text{ is among } k_\ell \text{ nearest neighbors} \\ & \text{of } \boldsymbol{x}_j, \\ & \text{or } \boldsymbol{x}_j \text{ is among } k_\ell \text{ nearest neighbors} \\ & \text{of } \boldsymbol{x}_i, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $k(\cdot, \cdot)$ represents the similarity between each pair of samples, and the standard measures thereof include heat kernel (Gaussian kernel), inner product, and euclidean distance. Note that in supervised cases, prior class information can be reflected to guide the graph construction [31]. Let $\boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$ be the normalized global and local Laplacian matrices, respectively, as

$$\begin{aligned} \boldsymbol{L}_g &= \boldsymbol{I} - \boldsymbol{D}_g^{-\frac{1}{2}}\boldsymbol{K}_g\boldsymbol{D}_g^{-\frac{1}{2}}, \\ \boldsymbol{L}_\ell &= \boldsymbol{I} - \boldsymbol{D}_\ell^{-\frac{1}{2}}\boldsymbol{K}_\ell\boldsymbol{D}_\ell^{-\frac{1}{2}}, \end{aligned}$$

where $\boldsymbol{D}_g$ and $\boldsymbol{D}_\ell$ are diagonal matrices such that $[\boldsymbol{D}_g]_{ii} = \sum_j [\boldsymbol{K}_g]_{ji}$ and $[\boldsymbol{D}_\ell]_{ii} = \sum_j [\boldsymbol{K}_\ell]_{ji}$. Then, we seek to find the projection matrix $\boldsymbol{W}$ that maximizes the following criterion:

$$J_{LLDA}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_g - 2\boldsymbol{S}_\ell)\boldsymbol{W}\big), \tag{5}$$

where $\boldsymbol{S}_g$ and $\boldsymbol{S}_\ell$ are the global and local scatter matrices defined as

$$\begin{aligned} \boldsymbol{S}_g &= \frac{1}{n}\boldsymbol{X}\boldsymbol{L}_g\boldsymbol{X}^T, \\ \boldsymbol{S}_\ell &= \frac{1}{n}\boldsymbol{X}\boldsymbol{L}_\ell\boldsymbol{X}^T. \end{aligned}$$

Note that the reason for using the normalized graph Laplacians is that the criterion (5) without normalization may be affected by the scale of the similarity measure or by the choice of the number of nearest neighbors, since (5) is defined as the difference rather than the ratio of the global scatter to the local scatter. Also, the use of normalized graph Laplacian is known to be effective in spectral clustering (e.g., [23]).

It is easy to check that, when we set $[\boldsymbol{K}_g]_{ij} = 1/n$ for all i, j; $[\boldsymbol{K}_\ell]_{ij} = 1/n_k$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are both in the kth class, and 0 otherwise, $\boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$, respectively, become $\boldsymbol{I} - \boldsymbol{W}_g$ and $\boldsymbol{I} - \boldsymbol{W}_\ell$; hence, (5) includes the MMC (2) as a special case.

In general, (5) does not require class information, thus can be used in an unsupervised manner. The construction of the local scatter matrix is based on the assumption that, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close, they are likely to belong to the same cluster. Under the condition that class labels are unavailable, we cannot explicitly consider the separability of different clusters, which is represented by the between-class scatter in classical LDA and the MMC. In the objective function (5), it is implicitly represented by the difference between the global scatter and the local scatter. Therefore, discriminative features can be extracted even in unsupervised scenarios. In this paper, we refer to unsupervised LLDA simply as LLDA.

3.2 Efficient Algorithm for LLDA

Similarly to the case of (2), the projection matrix $\boldsymbol{W}$ that maximizes (5) subject to the orthogonality constraint that $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$ can be found as the eigenvectors of $\boldsymbol{S}_g - 2\boldsymbol{S}_\ell$ corresponding to the largest eigenvalues. When p, the number of features, is very large as in microarray data, however, it is computationally demanding to directly perform the eigenvalue decomposition (EVD) of $\boldsymbol{S}_g - 2\boldsymbol{S}_\ell$, which is of size $p \times p$. In [26], two approaches for computing LLDA have been presented. The first one directly computes the eigenvalues and eigenvectors, hence demands expensive computational costs. The other approach achieves this via the spectral decomposition of Laplacian matrix, but it still needs to compute the eigenvalues and eigenvectors of a $p \times p$ matrix. Even worse, the eigenvectors corresponding to the nonpositive eigenvalues are discarded in the process of computing $\boldsymbol{W}$, thus it does not provide the exact solution to the maximization problem and may result in losing discriminatory information.

Here, we propose a novel algorithm for computing LLDA, which is particularly efficient when the feature size is much larger than the sample size, i.e., $p \gg n$, as is often the case with microarray data. The proposed algorithm is based on the following theorem (see Appendix A for the proof).

Theorem 1. *Let $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ be the reduced SVD [11] of $\boldsymbol{X} \in \mathbb{R}^{p\times n}$, where $\boldsymbol{P} \in \mathbb{R}^{p\times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ are orthonormal matrices and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ is a diagonal matrix. Further, let $\boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{V}^T$ be the EVD of a symmetric matrix $\boldsymbol{\Lambda}\boldsymbol{Q}^T(\boldsymbol{L}_g - 2\boldsymbol{L}_\ell)\boldsymbol{Q}\boldsymbol{\Lambda}$, where $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ is an orthonormal matrix and $\boldsymbol{\Delta} \in \mathbb{R}^{n\times n}$ is a diagonal matrix. Then, the projection matrix $\boldsymbol{W}$ of LLDA is constituted by the eigenvectors in $\boldsymbol{PV}$ corresponding to the largest eigenvalues in $\boldsymbol{\Delta}$.*

It is important to note that the main computation of the algorithm consists of the SVD of a $p \times n$ matrix and the EVD of an $n \times n$ matrix. Thus, it is very efficient in the case of $p \gg n$ (see also Appendix B for the computational complexity). The previous study [24] first removed the null space of the total scatter matrix via the SVD, thereby reducing the

dimensionality of the data to $n-1$, and then applied the MMC in the reduced space, where the rank of the mean-subtracted matrix of $\boldsymbol{X}$ was implicitly assumed to be $n-1$. Although the computational complexity of the proposed algorithm is the same as that of the previous MMC algorithm, LLDA can be directly applied to multicollinear data, while the previous MMC needs to estimate the rank of the mean-subtracted matrix in such a degenerate case.

In this way, the graph Laplacian representation of the MMC enables both the extension to unsupervised LLDA and the efficiency of the algorithm.

4 LLDA-RFE: Feature Selection Based on LLDA

The proposed algorithm for LLDA can be used in both supervised and unsupervised cases to extract discriminant features from high-dimensional data often encountered in, e.g., face recognition [16], [18], [31], [32], text categorization [5], [32], and microarray cancer classification [18], [32], [33]. In the context of microarray data analysis, the features so extracted correspond to *metagenes*, linear combinations of multiple genes, but we are rather interested in identifying discriminative genes themselves.

To this end, the previous study [24] proposed to combine the MMC with recursive feature elimination (RFE). The MMC-RFE algorithm recursively removes features with the smallest absolute values of the discriminant vectors of the MMC. The RFE approach has recently proven to be effective with regression [19], [34] as well as with support vector machine (SVM) [13]. In the present study, we propose an unsupervised recursive feature selection method using the discriminant vectors of LLDA to identify features that potentially reveal clusters in the samples. The proposed LLDA-RFE algorithm has the same feature elimination process as the MMC-RFE algorithm in [24], but LLDA-RFE does not involve dimension reduction unlike MMC-RFE. Another difference consists of the feature weighting scheme as described below.

While the number of discriminant vectors extracted by classical LDA is limited to at most $c-1$, the MMC and LLDA are capable of extracting more than $c-1$ discriminant vectors. It can also be shown that the maximum value of $J_{LLDA}(\boldsymbol{W})$ with the obtained discriminant vectors is equal to the sum of the corresponding eigenvalues. Because the eigenvalues reflect the discrimination ability, we use only the discriminant vectors corresponding to the positive eigenvalues to calculate the weight of each feature. Let $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n$ be the eigenvalues in $\boldsymbol{\Delta}$. Then, we define the weight of feature j as the sum of the absolute values of d discriminant vectors in $\boldsymbol{W}$, i.e., $\sum_{i=1}^{d} \sqrt{\delta_i}|[\boldsymbol{W}]_{ji}|$, where d is the number of positive eigenvalues. Here, the discriminant vectors are weighted by the magnitude of the corresponding eigenvalues.

Our proposed algorithm, LLDA-RFE, can be summarized as follows, and the computational complexity is given in Appendix B.

Algorithm: LLDA-RFE

Input: sample matrix: $\boldsymbol{X} \in \mathbb{R}^{p \times n}$
k_ℓ: the number of nearest neighbors
Output: r top-ranked features

0. Set $q \leftarrow p$;
Repeat the following steps until $q = r$
1. Construct the complete and k_ℓ-nearest neighbor graphs on $\boldsymbol{X}$ and compute $\boldsymbol{K}_g$, $\boldsymbol{K}_\ell$, $\boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$;
2. Perform the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^T$;
3. Compute $\boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{Q}^T(\boldsymbol{L}_g - 2\boldsymbol{L}_\ell)\boldsymbol{Q}\boldsymbol{\Lambda}$;
4. Perform the EVD of $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{V}^T$;
5. Set $\boldsymbol{W}$ to the eigenvectors in $\boldsymbol{PV}$ corresponding to the positive eigenvalues in $\boldsymbol{\Delta}$;
6. Remove the jth feature with the smallest weight of $\sum_{i=1}^{d} \sqrt{\delta_i}|[\boldsymbol{W}]_{ji}|$;
7. Set $q \leftarrow q-1$, form $\boldsymbol{X}$ and go to step 1.

5 Related Work on Unsupervised Feature Selection

Data variance is one of the most common criteria for unsupervised feature selection, and often used as a baseline method for comparison [15], [28]. Although variance ranking can be useful for selecting features that show large variations across all samples, it is not suited for selecting ones that contribute to characterizing different clusters in the samples. Hastie et al. [14] developed gene shaving to select informative genes from microarray data. Gene shaving iteratively removes genes having the lowest correlation with the leading principal component. Because the principal components are found so that they capture the directions of maximum variance in the data, gene shaving is also unsuitable for identifying genes that reveal different clusters. The assumption that discriminative genes exhibit large variance is not necessarily valid particularly for noisy microarray data, due to the large number of irrelevant genes. Indeed, recent studies [28], [30] have shown that variance ranking, principal component analysis, and gene shaving are not effective for yielding distinctive patterns between different classes of samples.

The latest and probably more effective unsupervised methods are Laplacian score [15], the $Q-\alpha$ algorithm [30], and SVD-entropy [28]. Among these, Laplacian score and SVD-entropy are employed for comparison in this study. In the following, we give a brief overview of the two methods. The $Q-\alpha$ algorithm is not included here due to expensive computational costs when applied to a data set with several thousand features.

5.1 Laplacian Score

The idea of Laplacian score is to evaluate each feature by its locality preserving power, showing similarity in spirit to Locality Preserving Projection [16].

Let $\boldsymbol{f}_r = (f_{r1}, \ldots, f_{rn})^T$, $r = 1, \ldots, p$, denote the rth feature for n samples. First, we construct a nearest neighbor graph in the same way as for the LLDA-RFE algorithm. Then, we compute the weight matrix $\boldsymbol{K}$, the diagonal matrix $[\boldsymbol{D}]_{ii} = \sum_j [\boldsymbol{K}]_{ji}$, and the graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$. Finally, the Laplacian score L_r of the rth feature is computed as

$$L_r = \frac{\tilde{\boldsymbol{f}}_r^T \boldsymbol{L} \tilde{\boldsymbol{f}}_r}{\tilde{\boldsymbol{f}}_r^T \boldsymbol{D} \tilde{\boldsymbol{f}}_r},$$

where

$$\tilde{\boldsymbol{f}}_r = \boldsymbol{f}_r - \frac{\boldsymbol{f}_r^T \boldsymbol{D}\boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{D}\boldsymbol{e}}\boldsymbol{e}.$$

According to the definition, a good feature should have a small value of the Laplacian score (see [15] for the detail). Thus, top-ranked features in this case are those with the smallest values of L_r. It is worth noting that Fisher score can be related to Laplacian score as shown in [15].

5.2 SVD-Entropy

Let us assume that $p > n$ for a given sample matrix $\boldsymbol{X} \in \mathbb{R}^{p \times n}$. Denoting by s_j the singular values of $\boldsymbol{X}$, an SVD-based entropy is defined as [2]

$$E = -\frac{1}{\log(n_r)} \sum_{j=1}^{n_r} V_j \log(V_j),$$

where

$$V_j = s_j^2 \Big/ \sum_{k=1}^{n_r} s_k^2.$$

Here, $n_r \leq n$ is the number of positive singular values, which is equal to the rank of $\boldsymbol{X}$. Then, the contribution of the ith feature to the entropy is defined as [28]

$$CE_i = E\big(\boldsymbol{X}_{[p \times n]}\big) - E\big(\boldsymbol{X}_{[(p-1) \times n]}\big),$$

where $\boldsymbol{X}_{[(p-1) \times n]}$ denotes the sample matrix with the ith feature being removed.

Varshavsky et al. [28] have proposed three feature selection strategies based on SVD-entropy: simple ranking (SR), forward selection (FS), and backward elimination (BE). Although SVD-entropy-based BE is somewhat similar to LLDA-RFE in the feature elimination process, it has high computational complexity in the case of a large number of features, hence impractical to apply to microarray data sets. This is due to the fact that CE_i is calculated on a leave-one-out basis. Indeed, Varshavsky et al. did not use BE in their experiments on microarrays. Accordingly, we employ SR in this study; top-ranked features are those with the largest values of CE_i.

6 Experimental Results

6.1 Data Sets and Preprocessing

In the experiments, we used seven public data sets of cancer microarrays. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, the different methods were compared primarily using binary-class data sets: ALL versus AML for Leukemia [10], normal versus tumor for Colon cancer [1], outcome prediction on Medulloblastoma [25], Breast cancer [27], and Lung adenocarcinoma [4]. In addition, we used multiclass data sets on MLL [3] and SRBCT [17] to further assess their performances. The characteristics of these data sets are summarized in Table 1, and the details are given below:

- Leukemia [10]: This Affymetrix high-density oligonucleotide array data set contains 38 samples from two classes of leukemia: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML). The data set is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- Colon cancer [1]: This Affymetrix high-density oligonucleotide array data set contains 62 samples from two classes of colon-cancer patients: 40 normal healthy samples and 22 tumor samples. The data set is publicly available at http://microarray.princeton.edu/oncology/affydata/index.html.
- Medulloblastoma data set [25]: This Affymetrix high-density oligonucleotide array data set contains 60 samples from two classes on patient survival with medulloblastoma: 21 treatment failures and 39 survivors. The data set is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- Breast cancer [27]: This cDNA microarray data set contains 76 samples from two classes on five-year metastasis-free survival: 33 poor prognosis and 43 good prognosis. The data set is publicly available at http://www.rii.com/publications/2002/vantveer.html.
- Lung adenocarcinoma [4]: This Affymetrix high-density oligonucleotide array data set contains 86 samples from two classes on survival: an event of death for 34 and alive for 52. The data set is publicly available at http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html.
- MLL [3]: This Affymetrix high-density oligonucleotide array data set contains 57 samples from three classes of leukemia: 20 acute lymphoblastic leukemia (ALL), 17 mixed-lineage leukemia (MLL), and 20 acute myelogenous leukemia (AML). The data set is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- SRBCT [17]: This cDNA microarray data set contains 63 samples from four classes of small round blue-cell tumors of childhood (SRBCT): 23 Ewing family of tumors, 20 rhabdomyosarcoma, 12 neuroblastoma, and 8 non-Hodgkin lymphoma. The data set is publicly available at http://research.nhgri.nih.gov/microarray/Supplement/.

TABLE 1
Characteristics of the Data Sets Used in This Study

Dataset	# samples	# classes	# genes
Leukemia	38	2	7129
Colon cancer	62	2	2000
Medulloblastoma	60	2	7129
Breast cancer	76	2	4918
Lung adenocarcinoma	86	2	7129
MLL	57	3	12582
SRBCT	63	4	2308

For the Leukemia, Medulloblastoma, Lung adenocarcinoma, and MLL data sets, expression values were first thresholded with a floor of 100 and a ceiling of 16,000, followed by a base 10 logarithmic transform. Then, each sample was standardized to zero mean and unit variance across genes. For the Colon cancer data set, after a base 10 logarithmic transform, each sample was standardized. For

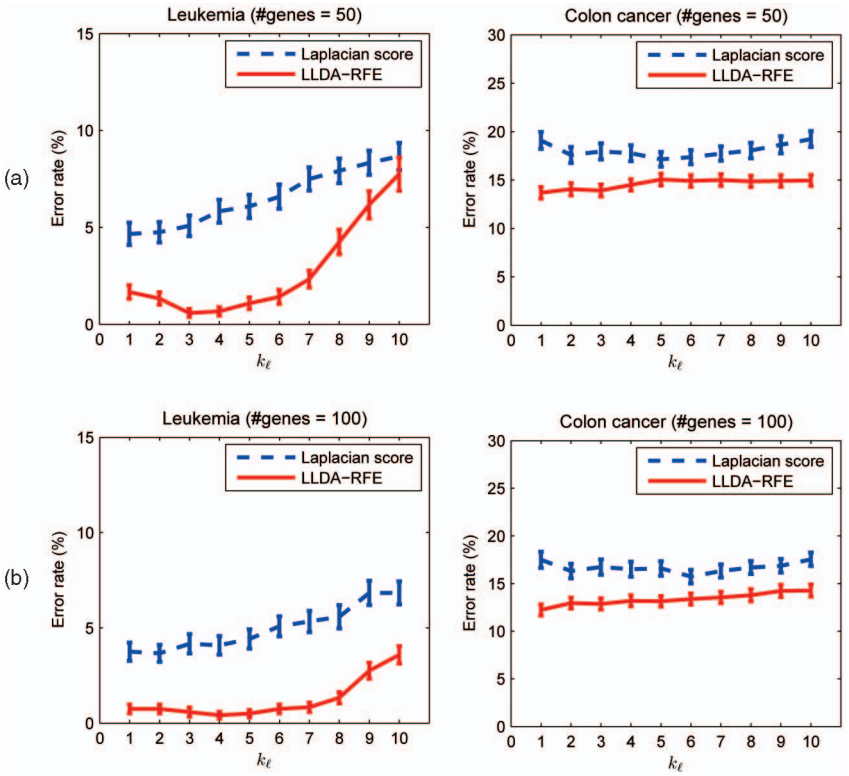

Fig. 1. Comparison of error rates between Laplacian score and LLDA-RFE using different values of k_ℓ for Leukemia and Colon cancer. The error rates are shown for (a) $\#\text{genes} = 50$ and (b) 100, respectively.

the Breast cancer data set, after the filtering of genes following [27], each sample was standardized. For the SRBCT data set, the expression profiles already preprocessed following [17] were used.

6.2 Performance Evaluation and Experimental Settings

We compare the performance of LLDA-RFE with those of Laplacian score and SVD-entropy, two state-of-the-art unsupervised feature selection methods. The performances of the unsupervised methods are evaluated by their capability of identifying discriminative genes without using class information. Varshavsky et al. [28] employed the Jaccard score of clustering algorithms such as K-means, showing how clusters can be discovered by using a smaller number of genes selected from several thousand or more genes in the same samples. Wolf and Shashua [30] measured the performance by the classification accuracy of a linear SVM classifier using leave-one-out cross-validation; gene selection was performed in an unsupervised setting, but classification in a supervised setting using the selected genes only. Because we also compare the performance between LLDA-RFE and Fisher score [8], [15], a supervised gene selection method, we employ the nearest mean classifier (NMC) and measure the performances by its classification accuracy. It is known that NMC is highly effective for cancer classification despite its simplicity [29]. Note that since Fisher score is supervised, it is generally expected to perform better than unsupervised methods.

We assessed the performance of each gene selection method with NMC by repeated random splitting as in [22]; the samples were partitioned randomly in a class proportional manner into a training set consisting of two-thirds of the whole samples and a test set consisting of the held-out one-third of the samples. To avoid selection bias, gene selection was performed using only the training set, and the classification error rate of the learnt classifier was obtained using the test set. This splitting was repeated 100 times. The error rates averaged over the 100 runs and the corresponding standard error rates are reported here.

To save computational time of RFE, we removed half of the genes until less than 500, and then a single gene at a time. For the computation of the weight matrix of Laplacian score and the similarity matrices of LLDA-RFE, we used the euclidean distance for nearest neighbor search and a simple 0-1 weighting as the similarity measure, i.e., $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are connected, and 0 otherwise.

6.3 Results and Discussion

6.3.1 *Effect of* k_ℓ

We first compare the performance between LLDA-RFE and Laplacian score by varying the number of nearest neighbors, on the binary-class data sets: Leukemia, Colon cancer, Medulloblastoma, Breast cancer, and Lung adenocarcinoma. Figs. 1 and 2 show the average error and standard error rates of NMC for $k_\ell = 1, 2, \ldots, 10$. For LLDA-RFE, k_ℓ was fixed to the same value during gene elimination. The numbers of genes selected and used for classification were 50 and 100.

It is clear that LLDA-RFE consistently achieves better performance than Laplacian score for all the data sets. This can be attributed to the difference that while Laplacian score is univariate, LLDA-RFE is multivariate and gene subsets are refined by the recursive elimination.

It may be difficult to set an appropriate value of k_ℓ in fully unsupervised settings because we cannot rely on

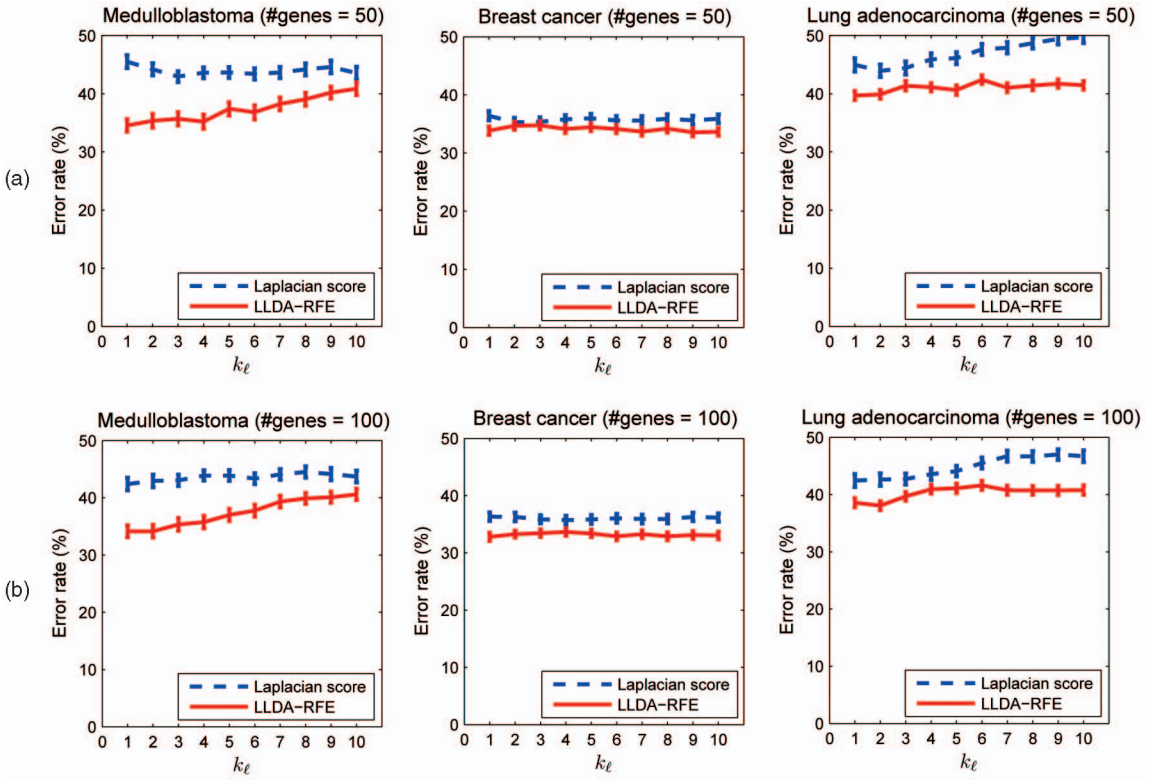

Fig. 2. Comparison of error rates between Laplacian score and LLDA-RFE using different values of k_ℓ for Medulloblastoma, Breast cancer, and Lung adenocarcinoma. The error rates are shown for (a) $\#\text{genes} = 50$ and (b) 100, respectively.

cross-validation unless class labels are provided and the value can also be largely dependent on the sample size of each data set and on the potential number of clusters therein. Although an adaptive setting of the value might be preferable during gene elimination, our results suggest that $k_\ell = 1$ to 3 is a reasonable choice when applying LLDA-RFE to microarray data sets with small sample size.

6.3.2 Comparison on Binary-Class Data Sets

Table 2 shows the average and standard error rates of NMC with four gene selection methods on the binary-class data sets. Fig. 3 plots the average error rates as a function of the number of genes from 1 to 100. The number of nearest neighbors for Laplacian score was set as follows: $k_\ell = 2$ for Leukemia, $k_\ell = 6$ for Colon cancer, $k_\ell = 3$ for Medulloblastoma and Breast cancer, and $k_\ell = 1$ for Lung adenocarcinoma. For LLDA-RFE, $k_\ell = 3$ was used for Leukemia and $k_\ell = 1$ for the other data sets.

We can observe that LLDA-RFE outperforms Laplacian score for a wide range of gene sizes. In comparison with SVD-entropy, LLDA-RFE yields lower error rates for Leukemia, Medulloblastoma, and Breast cancer. Although SVD-entropy appears to be better for Colon cancer and Lung adenocarcinoma, LLDA-RFE consistently shows satisfactory performances for all the data sets. Also, note that LLDA-RFE performs better than Fisher score for Leukemia, Medulloblastoma, and Breast cancer, despite the fact that LLDA-RFE is fully unsupervised. However, this does not imply that unsupervised gene selection is preferred to supervised one for these data sets. In fact, Fisher score, which can be viewed as a supervised version of Laplacian score, improves the performance of Laplacian score by using class information. Likewise, we can expect further improvement when using LLDA-RFE in a supervised manner.

6.3.3 Comparison on Multiclass Data Sets

Table 3 shows the average and standard error rates for the MLL and SRBCT data sets. Fig. 4 plots the average error rates as a function of the number of genes from 1 to 100. For Laplacian score, $k_\ell = 1$ was used for both data sets, and for LLDA-RFE, $k_\ell = 4$ and 3 were used for MLL and SRBCT, respectively.

It can be seen that LLDA-RFE reaches smaller error rates with a smaller number of genes, showing superior performance to Laplacian score and SVD-entropy. Notably, LLDA-RFE achieves even better performance than Fisher score. These results indicate that LLDA-RFE can also be useful for filtering genes from microarray samples potentially comprising multiple clusters.

In summary, our comparison using several microarray data sets has demonstrated that LLDA-RFE is effective for identifying genes that contribute to characterizing different clusters in the samples. Although we used 0-1 weighting as the similarity measure, the performance could be improved by using other data-dependent similarity measures. Also, more discriminative features can be found by balancing the global and local scatters as in (3).

7 Conclusions

In this paper, we have proposed an unsupervised feature selection method based on LLDA. In particular, we have

TABLE 2
Comparison of Error Rates (in Percentages) on Binary-Class Data Sets

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
Leukemia				
20	4.9 ± 0.6	**2.6 ± 0.4**	9.6 ± 1.0	3.6 ± 0.5
50	3.9 ± 0.4	1.6 ± 0.4	4.8 ± 0.5	**0.6 ± 0.2**
100	2.9 ± 0.4	1.6 ± 0.4	3.7 ± 0.4	**0.6 ± 0.2**
Colon cancer				
20	**12.4 ± 0.6**	14.9 ± 0.6	18.2 ± 0.8	15.9 ± 0.6
50	13.0 ± 0.6	**11.5 ± 0.6**	17.4 ± 0.7	13.7 ± 0.6
100	12.8 ± 0.5	**11.7 ± 0.6**	15.7 ± 0.7	12.2 ± 0.6
Medulloblastoma				
20	38.8 ± 0.9	36.8 ± 1.1	43.7 ± 1.0	**34.1 ± 1.1**
50	38.7 ± 1.0	38.4 ± 1.0	43.0 ± 0.9	**34.6 ± 1.1**
100	38.5 ± 1.0	37.1 ± 1.0	43.1 ± 1.0	**34.2 ± 1.1**
Breast cancer				
20	35.2 ± 0.9	42.6 ± 0.9	35.1 ± 0.8	**33.3 ± 0.8**
50	36.0 ± 0.8	42.0 ± 0.8	35.4 ± 0.8	**33.8 ± 0.7**
100	36.3 ± 0.8	42.4 ± 0.8	35.9 ± 0.7	**32.8 ± 0.7**
Lung adenocarcinoma				
20	**37.8 ± 0.8**	40.5 ± 0.8	45.7 ± 1.2	42.6 ± 0.7
50	**36.3 ± 0.8**	40.0 ± 0.7	45.0 ± 1.1	39.7 ± 0.8
100	**35.1 ± 0.8**	38.3 ± 0.8	42.4 ± 1.1	38.6 ± 0.8

Best results in boldface.

TABLE 3
Comparison of Error Rates (in Percentages) on Multiclass Data Sets

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
MLL				
20	7.2 ± 0.5	26.9 ± 0.9	10.2 ± 0.8	**6.1 ± 0.5**
50	6.6 ± 0.5	8.1 ± 0.6	9.4 ± 0.6	**5.1 ± 0.5**
100	5.9 ± 0.5	4.8 ± 0.4	9.1 ± 0.6	**3.8 ± 0.4**
SRBCT				
20	**3.6 ± 0.5**	22.6 ± 1.0	17.4 ± 1.2	16.2 ± 1.0
50	**2.6 ± 0.4**	17.4 ± 0.8	13.4 ± 1.0	12.0 ± 0.7
100	**4.6 ± 0.4**	11.8 ± 0.7	11.3 ± 0.9	11.4 ± 0.7

Best results in boldface.

extended LLDA to unsupervised cases and proposed a novel algorithm for computing LLDA, which is efficient for high-dimensional and small sample size data. The LLDA-RFE algorithm was then applied to several microarray data sets to identify discriminative genes without using class labels.

Our comparison with other state-of-the-art unsupervised feature selection methods and with a supervised method has demonstrated the feasibility and effectiveness of the proposed algorithm. LLDA-RFE is capable of identifying discriminative features that contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

A possible application of interest is the use of LLDA-RFE in semisupervised scenarios: When labels are partially given, we construct a graph such that samples from the same class are always connected, while those from different classes disconnected, and those with no class labels are adaptively connected or disconnected depending on the

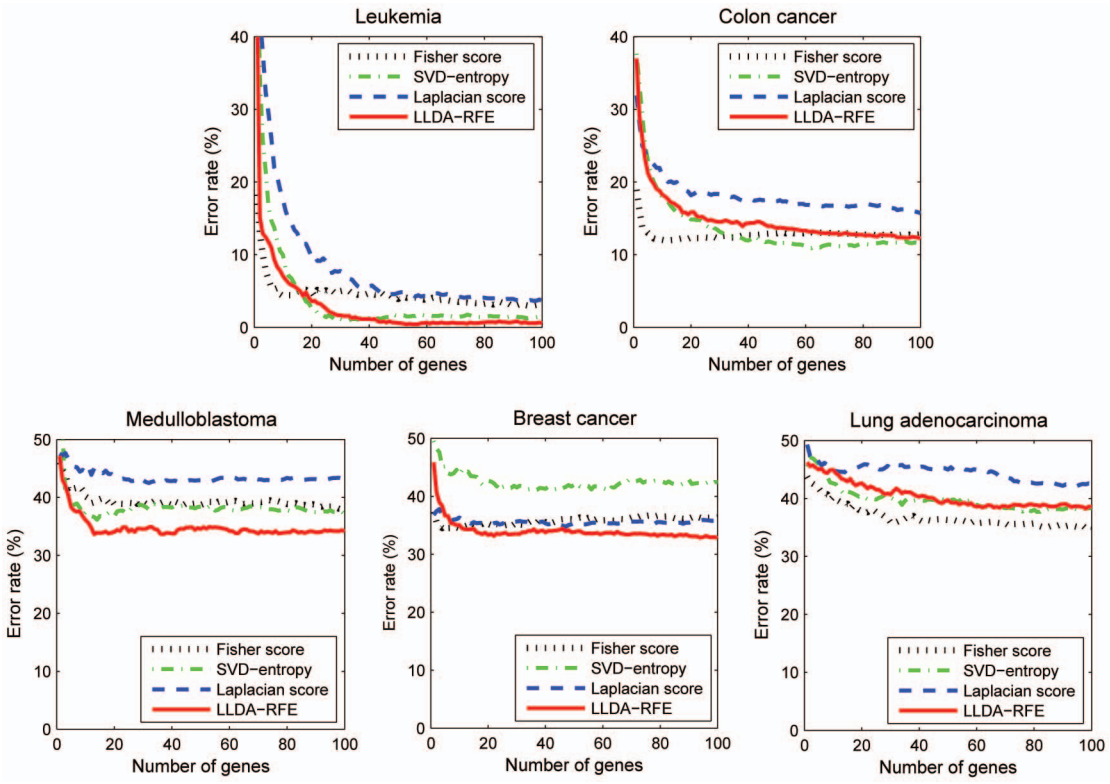

Fig. 3. The average error rates for Fisher score, SVD-entropy, Laplacian score, and LLDA-RFE on binary-class data sets.

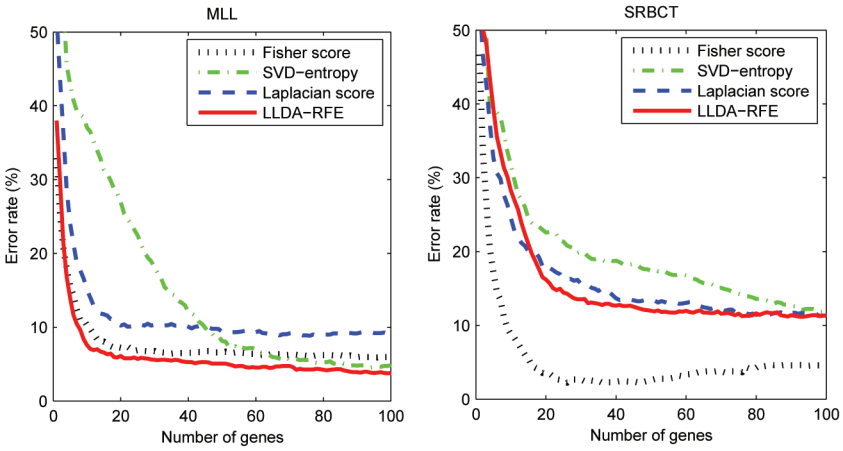

Fig. 4. The average error rates for Fisher score, SVD-entropy, Laplacian score, and LLDA-RFE on multiclass data sets.

nearest neighbors. We are currently exploring the applicability of LLDA-RFE as a semisupervised learning method.

Appendix A

Proof of Theorem 1

It is straightforward to verify that $S_g - 2S_\ell$ can be decomposed as follows:

$$\begin{aligned} S_g - 2S_\ell &= \frac{1}{n} X(L_g - 2L_\ell)X^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell)(P\Lambda Q^T)^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell) Q\Lambda P^T \\ &= \frac{1}{n} PV\Delta V^T P^T \\ &= \frac{1}{n} (PV)\Delta (PV)^T. \end{aligned}$$

Further, from the orthonormality of P and V

$$(PV)^T(PV) = V^T P^T PV = V^T V = I.$$

Thus, PV is the orthonormal matrix consisting of the eigenvectors of $S_g - 2S_\ell$. The projection matrix W of LLDA with the orthogonality constraint is constituted by the eigenvectors of $S_g - 2S_\ell$ corresponding to the largest eigenvalues, hence the theorem holds. □

Appendix B

Computational Complexity of the LLDA-RFE Algorithm

We analyze the time complexity of LLDA-RFE. The main computation in step 1 consists of k_ℓ-nearest neighbor search. The time complexity depends on the search algorithm employed, but when n is small, it does not affect the overall time of LLDA-RFE, hence omitted from this analysis. Step 2 takes $O(pn^2)$ time for the reduced SVD [11]. Steps 3 and 4 take $O(n^3)$ time for the matrix multiplications and for the EVD, respectively. Step 5 takes $O(pnd)$ time for computing PV, where $d(< n)$ is the number of positive eigenvalues. Step 6 takes $O(pd)$ time for calculating weights and $O(p)$ time for finding the smallest. Thus, in the case of $p \gg n$, the total time complexity for a single iteration is $O(pn^2)$. Overall, the LLDA-RFE algorithm takes $O((p^2 - r^2)n^2)$ time when recursively eliminating one feature at each iteration until the number of features reaches r. To alleviate the time complexity of RFE in the case of large p, a subset of features is often eliminated at a time.

Acknowledgments

This work was supported by a grant from the 21st Century COE program "Knowledge Information Infrastructure for Genome Science." A part of this work was done while Satoshi Niijima attended the Graduate School of Systems Life Sciences at Kyushu University, and also supported in part by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (to Professor S. Kuhara).

References

[1] 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 Sciences USA*, vol. 96, pp. 6745-6750, 1999.

[2] O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proc. Nat'l Academy of Sciences USA*, vol. 97, pp. 10101-10106, 2000.

[3] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia," *Nature Genetics*, vol. 30, pp. 41-47, 2002.

[4] D.G. Beer, S.L.R. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M.G. Taylor, M.D. Iannettoni, M.B. Orringer, and S. Hanash, "Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816-824, 2002.

[5] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 12, pp. 1624-1637, Dec. 2005.

[6] F.R.K. Chung, *Spectral Graph Theory, Regional Conf. Series in Math.*, no. 92, Am. Math. Soc., 1997.

[7] C.H.Q. Ding, "Unsupervised Feature Selection via Two-Way Ordering in Gene Expression Analysis," *Bioinformatics,* vol. 19, no. 10, pp. 1259-1266, 2003.

[8] S. Dudoit, J. Fridlyand, and T. 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] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Academic Press, 1990.

[10] 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,* vol. 286, pp. 531-537, 1999.

[11] G.H. Golub and C.F. Van Loan, *Matrix Computations,* third ed. Johns Hopkins Univ. Press, 1996.

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[14] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P.O. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology,* vol. 1, no. 2, research0003, 2000.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," *Advances in Neural Information Processing Systems 18,* Y. Weiss, B. Schölkopf, and J. Platt, eds., pp. 507-514, MIT Press, 2006.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 328-340, Mar. 2005.

[17] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine,* vol. 7, no. 6, pp. 673-679, 2001.

[18] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 157-165, 2006.

[19] F. Li and Y. Yang, "Analysis of Recursive Gene Selection Approaches from Microarray Data," *Bioinformatics,* vol. 21, no. 19, pp. 3741-3747, 2005.

[20] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face Recognition Using Kernel Scatter-Difference-Based Discriminant Analysis," *IEEE Trans. Neural Networks,* vol. 17, no. 4, pp. 1081-1085, 2006.

[21] M. Loog, "On an Alternative Formulation of the Fisher Criterion That Overcomes the Small Sample Problem," *Pattern Recognition,* vol. 40, pp. 1753-1755, 2007.

[22] S. Michiels, S. Koscielny, and C. Hill, "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy," *Lancet,* vol. 365, pp. 488-492, 2005.

[23] A.Y. Ng, M.I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Advances in Neural Information Processing Systems 14,* T. Dietterich, S. Becker, and Z. Ghahramani, eds., pp. 849-856, MIT Press, 2002.

[24] S. Niijima and S. Kuhara, "Recursive Gene Selection Based on Maximum Margin Criterion: A Comparison with SVM-RFE," *BMC Bioinformatics,* vol. 7, 543, 2006.

[25] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H 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. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, "Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression," *Nature,* vol. 415, pp. 436-442, 2002.

[26] H. Tang, T. Fang, and P.-F. Shi, "Laplacian Linear Discriminant Analysis," *Pattern Recognition,* vol. 39, pp. 136-139, 2006.

[27] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature,* vol. 415, pp. 530-536, 2002.

[28] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *Bioinformatics,* vol. 22, no. 14, pp. e507-e513, 2006.

[29] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veenman, H. Dai, Y.D. He, and L.J. van't Veer, "A Protocol for Building and Evaluating Predictors of Disease State Based on Microarray Data," *Bioinformatics,* vol. 21, no. 19, pp. 3755-3762, 2005.

[30] L. Wolf and A. Shashua, "Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach," *J. Machine Learning Research,* vol. 6, pp. 1855-1887, 2005.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 1, pp. 40-51, Jan. 2007.

[32] J. Ye, "Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems," *J. Machine Learning Research,* vol. 6, pp. 483-502, 2005.

[33] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 1, no. 4, pp. 181-190, Oct.-Dec. 2004.

[34] J. Zhu and T. Hastie, "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics,* vol. 5, no. 3, pp. 427-443, 2004.

Satoshi Niijima received the MEng degree in information science and the PhD degree in systems life sciences from Kyushu University, Fukuoka, Japan, in 2000 and 2007, respectively. From 2000 to 2002, he was in the Department of Electrical Engineering, University of Tokyo, where he studied biomedical engineering. He is currently a postdoctoral fellow in the Graduate School of Pharmaceutical Sciences, Kyoto University. His research interests include machine learning, pattern recognition, data mining, bioinformatics, and chemoinformatics.

Yasushi Okuno received the PhD degree in pharmaceutical sciences from Kyoto University, Kyoto, Japan, in 2000. He is currently an associate professor in the Graduate School of Pharmaceutical Sciences, Kyoto University. His main research interests include bioinformatics, chemoinformatics, and integration of these two research fields.