Data & Knowledge Engineering 66 (2008) 338–347

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Cancer classification by gradient LDA technique using microarray gene expression data

Alok Sharma [a,b,*], Kuldip K. Paliwal [a]

[a] Signal Processing Lab, Griffith University, Brisbane, Australia
[b] University of the South Pacific, School of Engineering and Physics, Suva, Fiji

ARTICLE INFO

Article history:
Received 18 December 2007
Received in revised form 29 March 2008
Accepted 8 April 2008
Available online 22 April 2008

Keywords:
DNA microarray
Linear discriminant analysis (LDA)
Gradient LDA (GLDA)
Dimensionality reduction
Cancer classification
Feature selection
Feature extraction

ABSTRACT

Cancer classification is one of the major applications of the microarray technology. When standard machine learning techniques are applied for cancer classification, they face the small sample size (SSS) problem of gene expression data. The SSS problem is inherited from large dimensionality of the feature space (due to large number of genes) compared to the small number of samples available. In order to overcome the SSS problem, the dimensionality of the feature space is reduced either through feature selection or through feature extraction. Linear discriminant analysis (LDA) is a well-known technique for feature extraction-based dimensionality reduction. However, this technique cannot be applied for cancer classification because of the singularity of the within-class scatter matrix due to the SSS problem. In this paper, we use Gradient LDA technique which avoids the singularity problem associated with the within-class scatter matrix and shown its usefulness for cancer classification. The technique is applied on three gene expression datasets; namely, acute leukemia, small round blue-cell tumour (SRBCT) and lung adenocarcinoma. This technique achieves lower misclassification error as compared to several other previous techniques.

1. Introduction

The advent of microarray technology has enabled the researchers to rapidly measure the levels of thousands of genes expressed in a biological tissue sample in a single experiment [19]. One important application of this microarray technology is to classify the tissue samples using their gene expression profiles as one of the several types (or subtypes) of cancer. Compared with the standard histopathological tests, the gene expression profiles measured through microarray technology provide accurate, reliable and objective cancer classification.

The DNA microarray data for cancer classification consists of large number of genes (dimensions) compared to the number of samples or feature vectors. The high dimensionality of the feature space degrades the generalisation performance of the classifier and increases its computational complexity. This problem is popularly in known as the small sample size (SSS) problem in the literature [10]. It restricts direct application of conventional statistical and machine learning techniques for classification purposes. This situation, however, can be overcome by first reducing the dimensionality of feature space, followed by classification in the lower-dimensional feature space. Different methods used for dimensionality reduction can be grouped into two categories: feature selection methods and feature extraction methods. Feature selection methods retain

[*] Corresponding author. Address: University of the South Pacific, School of Engineering and Physics, Suva, Fiji. Tel.: +679 3232870; fax: +679 3231538.
E-mail addresses: sharma_al@usp.ac.fj (A. Sharma), K.Paliwal@griffith.edu.au (K.K. Paliwal).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.04.004

only a few useful features and discard others. Feature extraction methods construct a few features from the large number of original features through their linear (or nonlinear) combination. For the classification of the lower-dimensional feature vector, the Bayes decision rule provides the most optimal solution. But, since the amount of training data available for designing the classifier is limited and small, other classifiers (e.g., nearest centroid classifier, k-nearest neighbour classifier, etc.) are used for cancer classification. A number of papers have been reported in the past for the cancer classification task using the microarray data. We provide a brief description of a small sample of these papers to highlight different techniques used for dimensionality reduction and classification for this task.

Golub et al. [13] adopted gene selection criteria based on correlation of genes prior to the classification. The selected genes were utilized in weighted voting (WV) approach for cancer classification. Furey et al. [11] applied similar technique as of Golub et al. [13] for gene selection and demonstrated the use of support vector machine (SVM) for cancer classification. Dudoit et al. [8] compared the performance of different discrimination methods for classification of tumours. These methods included nearest neighbour (NN) classifier, linear discriminant analysis (LDA), diagonal discriminant analysis, quadratic classifiers and classification trees. They considered bagging [5] and boosting [9] approaches to select relevant genes, which were used in the classification. Nguyen and Rocke [18] proposed partial least square (PLS) method for human tumor classification. They used PLS and principal component analysis (PCA) for dimension reduction as well as quadratic discriminant analysis (QDA) and logistic discrimination (LD) for classification task. Guyon et al. [14] proposed a gene selection criterion utilizing SVM methods based on recursive feature elimination (RFE). Lee et al. [16] developed a hierarchical Bayesian (HB) model for variable gene selection. Instead of fixing the number of selected genes (dimensions), a prior distribution over it was assigned. Bee and Mallick [2] pointed out that this approach is sensitive toward the choice of some hyper-parameters. Consequently, they considered a multivariate Bayesian regression model and assigned *priors* that favour sparseness in terms of number of genes used. They introduced the use of different *priors* to promote different degree of sparseness using a two-level hierarchical Bayesian (2L-HB) model. Zhou et al. [28] proposed a Bayesian approach to gene selection and classification using logistic regression model [1]. They used Gibbs sampling and Markov chain Monte Carlo (MCMC) methods to discover important genes. Geman et al. [12] introduced top scoring pair (TSP), which is based on pairwise comparison between two gene expression levels. This TSP algorithm was extended by Tan et al. [24] to k-TSP, which uses k pairs of genes for classifying gene expression data. They investigated the performance of TSP and k-TSP for three different schemes namely one-vs-others scheme, one-vs-one scheme and hierarchical classification (HC) scheme. Yeung et al. [26] used Bayesian model averaging (BMA) to address multi-class cancer classification problem. A typical gene selection and classification procedure ignores model uncertainty and uses a single set of relevant genes to predict the class. On the other hand, BMA accounts for the uncertainty by averaging over multiple sets of potentially overlapping relevant genes. Tan et al. [25] addressed the small sample size problem with microarray data by proposing total principal component regression (TPCR). It can classify human tumors by extracting the lateral variable structure underlying microarray data from the augmented subspace of both independent variables and dependent variables. Zhang et al. [27] developed a type of regularization in SVM to identify important genes for cancer classification. Leng and Müller [17] used functional logistic regression tool based on functional principal components for classifying temporal gene expression data.

From this short survey, it is clear that most of the techniques described above employ feature (or gene) selection for dimensionality reduction. Since feature extraction method always give better performance than feature selection method [7] we will investigate techniques based on feature extraction methods in this paper. In the feature (or gene) selection methods, several genes are discarded. Only a few genes are retained based on some criterion function, which tries to rank genes for classification purpose. The genes that are discarded may contain crucial information for classification of cancer types. In addition, the choice of the number of genes to be selected in these papers is usually arbitrary. Theoretically it is difficult to identify the number of selected genes that provides optimal performance for the classification of cancerous tissues, though empirical arguments can be made to justify the number of genes to be selected.

In the present paper, we concentrate on the feature extraction methods for dimensionality reduction. Feature extraction methods can provide better classification performance over feature selection methods since in feature extraction the subspace is a linear combination of original feature space [7]. The two popular feature extraction methods for dimensionality reduction are principal component analysis (PCA) and linear discriminant analysis (LDA). The former method concentrates on the representation of data and is not very powerful in discriminating the cancer classes. The LDA method, on the other hand, is discriminative in nature and could help in classifying a tissue sample more accurately. In LDA, we attempt to maximize between-class scatter with respect to within-class scatter. This process requires solving generalized eigenvalue decomposition problem, which involves the computation of the inverse of within-class scatter matrix. Due to the high dimensionality of microarray data compared to the number of samples available, the scatter matrix becomes singular and its inverse computation is not feasible. Looking at the limitations of LDA, we adapted gradient LDA (GLDA) technique [21] which resolves this type of limitation. The GLDA technique utilizes gradient descent algorithm to do dimensionality reduction. Once the dimension is reduced through the GLDA algorithm, the k-nearest neighbour classifier with Euclidean distance measure is used to classify a tissue sample. Experiments on several microarray gene expression datasets using the GLDA technique show very encouraging results for cancer classification.

The paper is organized as follows. Section 2 describes the three datasets used in the experimentation, Section 3 illustrates the GLDA technique, Section 4 describes the experiments and results, and Section 5 presents our conclusions.

2. Description of datasets used in the experimentation

Three datasets are utilized in this work to show the effectiveness of the proposed algorithm. The datasets are acute Leukemia [13], SRBCT [15] and Lung Adenocarcinoma [3]. The description of the datasets is given as follows:

(1) Acute Leukemia dataset [13]: this dataset consists of DNA microarray gene expression data of human acute leukemias for cancer classification. Two types of acute leukemias data are provided for classification namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset is subdivided into 38 training samples and 34 test samples. The training set consists of 38 bone marrow samples (27 ALL and 11 AML), over 7129 probes. The testing set consists of 34 samples with 20 ALL and 14 AML, prepared under different experimental conditions. All the samples have 7129 dimensions and all are numeric.

(2) SRBCT dataset [15]: the small round blue-cell tumor dataset consists of 83 samples with each having 2308 genes. This is a four class classification problem. The tumors are Burkitt lymphoma (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The testing set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively.

(3) Lung Adenocarcinoma dataset [3]: this dataset consists of 96 samples each having 7129 genes. This is a three class classification problem. Out of 96 samples 86 are primary lung adenocarcinomas, including 67 stage I tumor and 19 stage III tumor. An addition of 10 non-neoplastic lung samples is provided. The dataset was not prearranged for training set and testing set, therefore to conduct the experimentation we subdivided the dataset into training set and testing set in the following manner. The training set is allocated 64 samples which includes first 44 samples of stage I tumor, first 13 samples of stage III tumor and first 7 samples of non-neoplastic class. The remaining 23 stage I tumor samples, 6 stage III tumor samples and 3 non-neoplastic lung samples are utilized for testing.

3. The GLDA technique for cancer classification

The GLDA technique avoids the SSS problem, typically encountered in cancer classification using microarray gene expression data. We begin this section with a brief description of the conventional LDA technique including its main limitations, the singularity issue due to the SSS problem. We then discuss some methods that have been used in the past to overcome the SSS problem. This is followed by a discussion of the GLDA technique.

3.1. Linear discriminant analysis

Linear discriminant analysis is a popular method for feature extraction and dimensionality reduction. It operates in a supervised mode. In order to describe this method, we define first the notation used. The matrix $\mathcal{X}$ denotes a d-dimensional set of n training samples in a c-class problem, $\Omega = \{\omega_i : i = 1, 2, \ldots, c\}$ denotes the finite set of c-class labels, where ω_i represents the ith class label. The set $\mathcal{X}$ is partitioned into c subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_c$ where each subset $\mathcal{X}_i$ belongs to class ω_i and consists of n_i number of samples such that:

$$n = \sum_{i=1}^{c} n_i$$

The samples of set $\mathcal{X}$ can be written as

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \text{ where } \mathbf{x}_j \in \mathbf{R}^d \ (d\text{-dimensional hyperplane})$$

$$\mathcal{X}_i \subset \mathcal{X} \text{ and } \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_c = \mathcal{X}$$

Given the sample set $\mathcal{X}$, the within-class scatter matrix (S_{W}) and the between-class scatter matrix (S_{B}) used in LDA can be defined as

$$S_{\mathrm{W}} = \sum_{j=1}^{c} \sum_{\mathbf{x} \in \mathcal{X}_j} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}j})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}j})^{\mathrm{T}} \tag{1}$$

$$\text{and} \quad S_{\mathrm{B}} = \sum_{j=1}^{c} n_j (\boldsymbol{\mu}_{\mathbf{x}j} - \boldsymbol{\mu}_{\mathbf{x}})(\boldsymbol{\mu}_{\mathbf{x}j} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{\mu}_{\mathbf{x}\,\mathrm{j}}$ is the centroid of $\mathcal{X}_j$ and $\boldsymbol{\mu}_{\mathbf{x}}$ is the centroid of $\mathcal{X}$. In LDA, the d-dimensional space is transformed to h-dimensional space (where $1 \leqslant h \leqslant c - 1$) in such a way that the Fisher's criterion [7]

$$J(\mathbf{W}) = \frac{|\mathbf{W}^{\mathrm{T}} S_{\mathrm{B}} \mathbf{W}|}{|\mathbf{W}^{\mathrm{T}} S_{\mathrm{W}} \mathbf{W}|} \tag{3}$$

is maximised, where $|\bullet|$ denotes the determinant. The projection is from d-dimensional space to h-dimensional space $\mathbf{W}:\mathbf{x} \to \mathbf{y}$ or $\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{x} \in \mathbf{R}^d$ and $\mathbf{y} \in \mathbf{R}^h$. $\mathbf{W}$ is a $d \times h$ matrix representing the LDA transformation and is the solution of the conventional eigenvalue problem

$$S_W^{-1}S_B\mathbf{w}_i = \lambda_i\mathbf{w}_i \tag{4}$$

where $\mathbf{w}_i$ are the column vectors of **W**. The desired h leading eigenvectors are selected for **W** that correspond to the largest eigenvalues (λ_i) in Eq. (4). Clearly, from Eq. (4) it is evident that the explicit solution for the orientation **W** can be found only when S_W is non-singular. In the cancer classification problem using DNA microarray gene expression data, the matrix S_W turns out to be singular due to the SSS problem which restricts the direct application of the conventional LDA technique.

Singularity problem can be avoided in a number of ways. For example, Dudoit et al. [8] have used feature selection process to reduce the number of features from d-dimensional space to a smaller p-dimensional space. Then LDA technique was applied on these p-dimensional features. But the results obtained by Dudoit et al. [8] were not very encouraging for the application of LDA technique. This is due to the fact that some features containing useful information crucial for classification could have been discarded during the feature selection process.

The SSS problem has also been addressed in the literature by using PCA technique prior to the application of LDA [23,4,20]. The PCA technique reduces the d-dimensional space to smaller dimensional space to make the within-class scatter matrix non-singular. The PCA technique is not optimized for finding discriminant features and therefore could discard some features crucial for classification purpose.

Some techniques that are based on the modified Fisher's criterion [6,22] have also been developed. These techniques do not optimise the Fisher's criterion function in one stage. As a result, they are not optimal.

On the other hand, the GLDA technique avoids the singularity issue associated with the within-class scatter matrix by optimising the Fisher's criterion. Thereby finding the leading eigenvectors for singular S_W is possible provided rank(S_W) $\geqslant h$ and rank(S_B) $\geqslant h$ for DNA microarray data based cancer classification problem.

The orientation **W** can be evaluated for singular S_W by using gradient descent algorithm based on Fisher's criterion. The **W** value that gives maximum $J(\mathbf{W})$ is the desired orientation. The maximization problem can also be transformed to the minimization problem by denoting $\widehat{J}(\mathbf{W}) = 1/J(\mathbf{W})$ and finding **W** for which $\widehat{J}(\mathbf{W})$ is minimum. The GDLA algorithm can be given as [21]

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha\frac{\partial\widehat{J}(\mathbf{W})}{\partial\mathbf{W}} \tag{5}$$

$$\mathbf{W} \leftarrow \text{Normalize each of the column vectors of } \mathbf{W} \text{ separately} \tag{6}$$

where $\frac{\partial\widehat{J}(\mathbf{W})}{\partial W} = 2\widehat{J}(\mathbf{W})[S_W\mathbf{W}(\mathbf{W}^{\mathrm{T}}S_W\mathbf{W})^{-1} - S_B\mathbf{W}(\mathbf{W}^{\mathrm{T}}S_B\mathbf{W})^{-1}]$ and $\alpha > 0$ is the learning rate parameter.

The GLDA algorithm computes **W** in an iterative manner. The equation (5) updates **W** in the direction of steepest descent and Eq. (6) normalizes each column of updated **W** separately. The iterative process of the algorithm is terminated when $J(\mathbf{W})$ converges. In the algorithm we initialize the orientation using a fixed value for **W** as follows:

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 1 & 1 & 0 & \cdots \\ 1 & 0 & 1 & \cdots \\ 1 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

4. Experiments and results

This section demonstrates the performance of the GLDA approach in comparison with many previously reported techniques on gene expression datasets. Three sets of microarray gene expression datasets for cancer classification are utilized namely acute leukemia, SRBCT and lung adenocarcinoma which are described in detail in Section 2. The training set of data is utilized for learning the parameters of the classifier, while the testing set of data is utilized to measure the misclassification error.

4.1. Training phase

The training phase involves the following steps:

(1) Given the DNA microarray gene expression data $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_j \in \mathbf{R}^d$, $j = 1, \ldots, n$, compute matrices S_W and S_B using Eqs. (1) and (2).

(2) Find the orientation **W** using the gradient LDA algorithm. This algorithm requires a proper value for learning rate parameter α. In our experiments, we have selected the value by investigating the learning curve (plot of $J(\mathbf{W})$ as a function of iteration number) and keeping convergence and stability in mind.

(3) Transform the data from original d-dimensional space to h-dimensional space using **W**, i.e., $\mathbf{y}_j = \mathbf{W}^{\mathrm{T}}\mathbf{x}_j$ where $\mathbf{y}_j \in \mathbf{R}^{\mathrm{h}}$ and $j = 1, \ldots, n$.

4.2. Testing phase

The testing or classification phase is described as follows:

(1) Given a test sample $\mathbf{x} \in \mathbf{R}^{\mathrm{d}}$. Project the test sample to h-dimensional space using the orientation $\mathbf{W}$; i.e., $\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$.
(2) Compute the Euclidean distance $\delta_j = \|\mathbf{y} - \mathbf{y}_j\|$ for $j = 1, 2, \ldots, n$.
(3) Find the k-nearest neighbours using δ_j values and assign the test vector to the class label represented most in the k-nearest neighbours.

4.3. An illustration using acute leukemia dataset

This section illustrates the GLDA technique in comparison with PCA technique using acute leukemia dataset. The dimensionality of feature vectors of acute leukemia dataset is 7129. It has two classes ($c = 2$) namely ALL and AML. The GLDA technique can reduce the dimensions from 7129 to h, where $1 \leqslant h \leqslant c - 1$. This turns out the maximum value of h to 1 as the

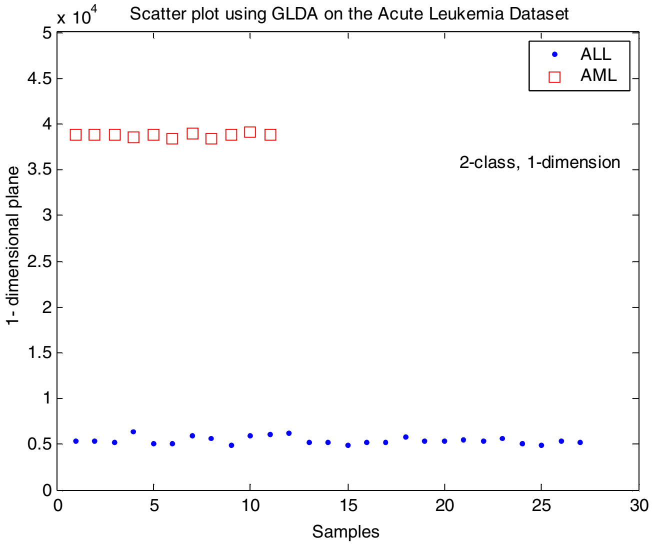

Fig. 1a. Scatter plot of training data of acute leukemia using GLDA technique.

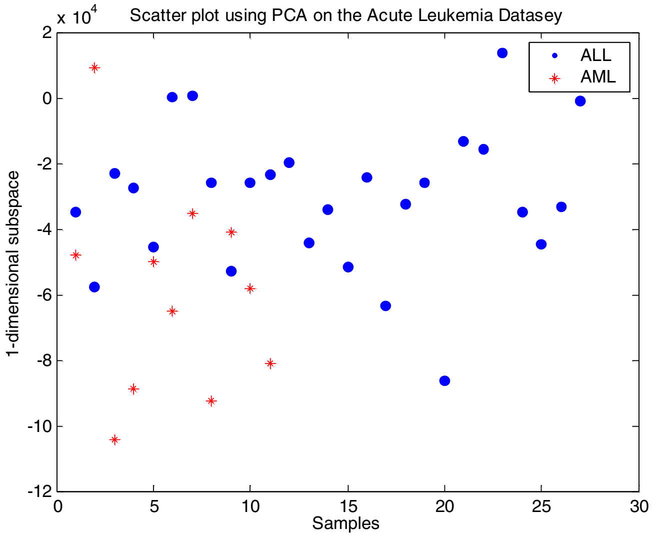

Fig. 1b. Scatter plot of training data of acute leukemia using PCA technique.

class is 2. The scatter plot of training data of acute leukemia using GLDA technique is shown in Fig. 1a. It can be seen that the training data of different classes are well separated which will give better classification performance. In order to compare the scatter plot of GLDA technique, we used scatter plot derived from PCA technique which is depicted in Fig. 1b. It is evident from the figure that the feature vectors of adjacent classes overlap with each other. That means the performance of PCA technique will not be very encouraging. This will produce high misclassification error. The classification accuracy using GLDA technique was evaluated 100% whereas the classification accuracy using PCA technique was calculated to be 64.71%. This simple example gives us an encouragement that GLDA technique may produce better classification performance when compared with other standard techniques.

4.4. Results

The experimentation was done using Matlab software. The classification results using GLDA technique on acute leukemia dataset are shown in Fig. 2. This dataset has samples from two cancer classes (ALL and AML). Thus, the GLDA technique is applied on a two-class problem and reduces the dimensionality from 7129-dimensional space to 1-dimensional subspace. The projection of ALL and AML samples on 1-dimensional subspace is depicted in Fig. 2a. The y-axis represents 1-dimensional subspace and x-axis represents the samples used during training and testing phases. Fig. 2b illustrates the plot showing the logarithm of Fisher's criterion ($J(\mathbf{W})$) as a function of iteration number for $\alpha = 0.1$ using the training data. This type of

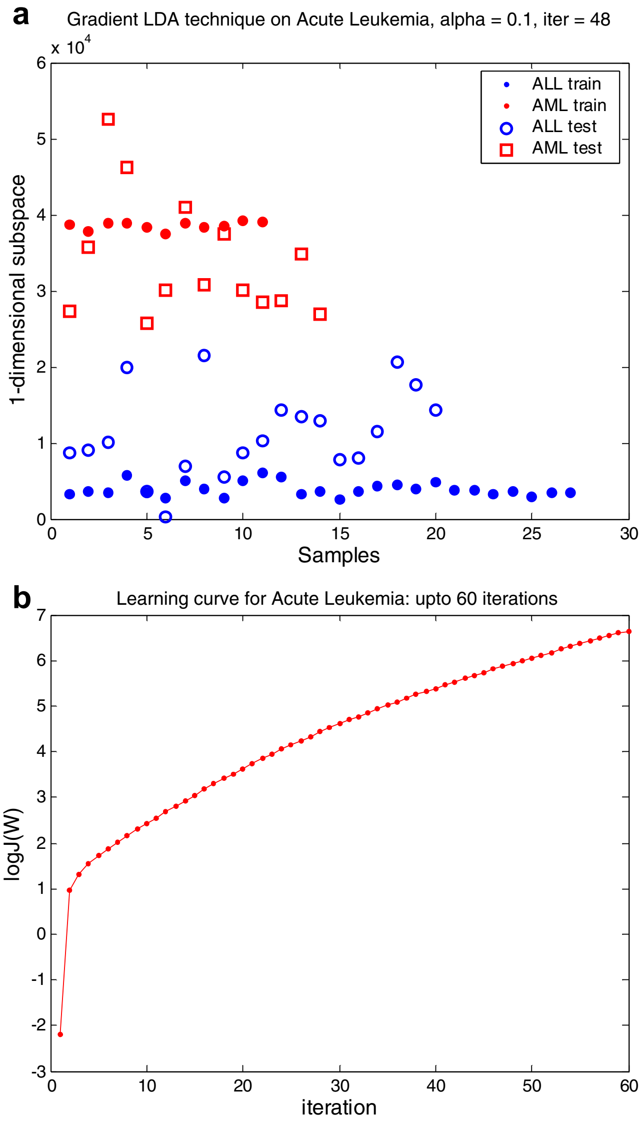

Fig. 2. (a) Classification using GLDA technique on acute leukemia dataset and (b) learning curve for acute leukemia dataset.

Table 1
Misclassification on acute leukemia dataset

Method	Number of misclassified samples
SVM-RFE [14]	0–10
QDA [18]	2–6
SVM [11]	2–4
LD [18]	1–4
WV [13]	2
BMA [26]	2
2L-HB [2]	2
GLDA	0

plot is commonly known in the literature as the learning curve. This figure shows that *J*(**W**) initially increases with an increase in the number of iterations and after some iterations, it saturates (converges). We say that the GLDA algorithm has converged when the relative increase in the successive iterations is about 1%. For the acute leukemia database the convergence was reached after 48 iterations. It can be observed from Fig. 2a that the training data of ALL and AML classes are well separated. It can also be noted that even the test data of ALL and AML classes are well separated and are close to their corresponding training data. The *k*-nearest neighbour classifier is used here to classify the reduced (1-dimensional) vector into

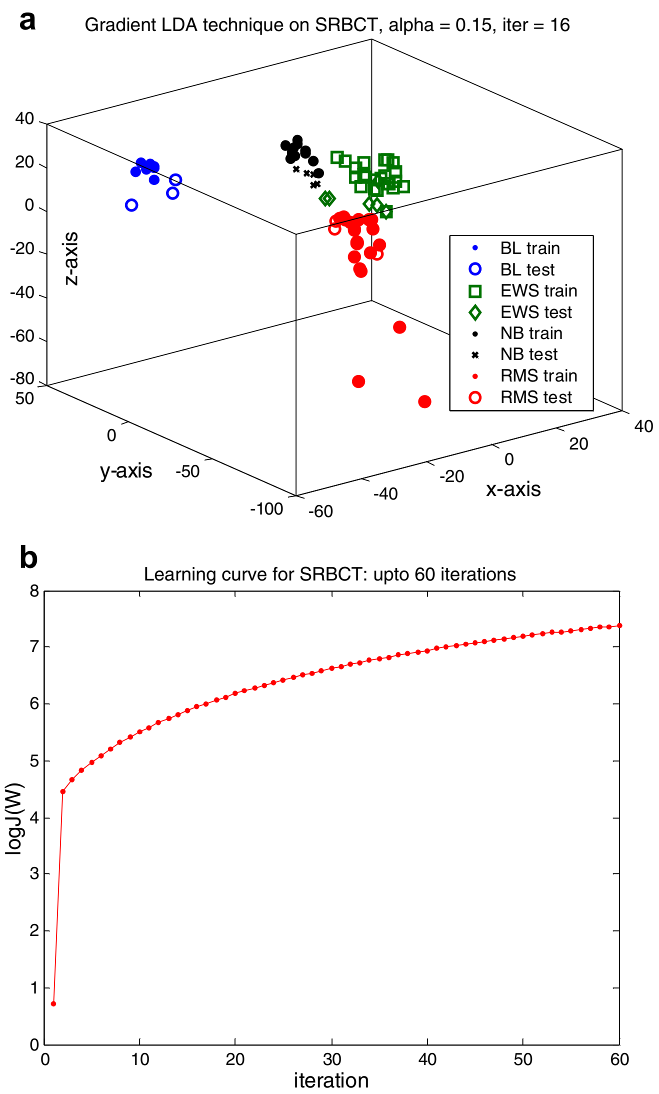

Fig. 3. (a) Classification using GLDA technique on SRBCT dataset and (b) learning curve for SRBCT dataset.

Table 2
Misclassification on SRBCT dataset

Method	Number of misclassified samples
HC-TSP [24]	1
HC-*k*-TSP [24]	0
TPCR [25]	0
GLDA	0

two cancer classes. The value of *k* (obtained by cross validation on training data) is found to be one. The number of misclassified samples by the GLDA and other techniques is shown in Table 1. It is evident that the GLDA technique is giving zero misclassification error which is better than all the other presented techniques in Table 1. It should also be noted that this zero misclassification error has been obtained using only 1-dimensional vector.

The classification results using GLDA technique on SRBCT dataset are shown in Fig. 3. The dataset has four types of tumor samples namely BL, EWS, NB and RMS. The GLDA technique is therefore applied on 4-class problem and reduces the dimensionality from 2308 to 3. The projection of samples from BL, EWS, NB and RMS classes on 3-dimensional subspace is depicted in Fig. 3a. It can be observed from Fig. 3a that the training samples of different classes are well separated from each other on

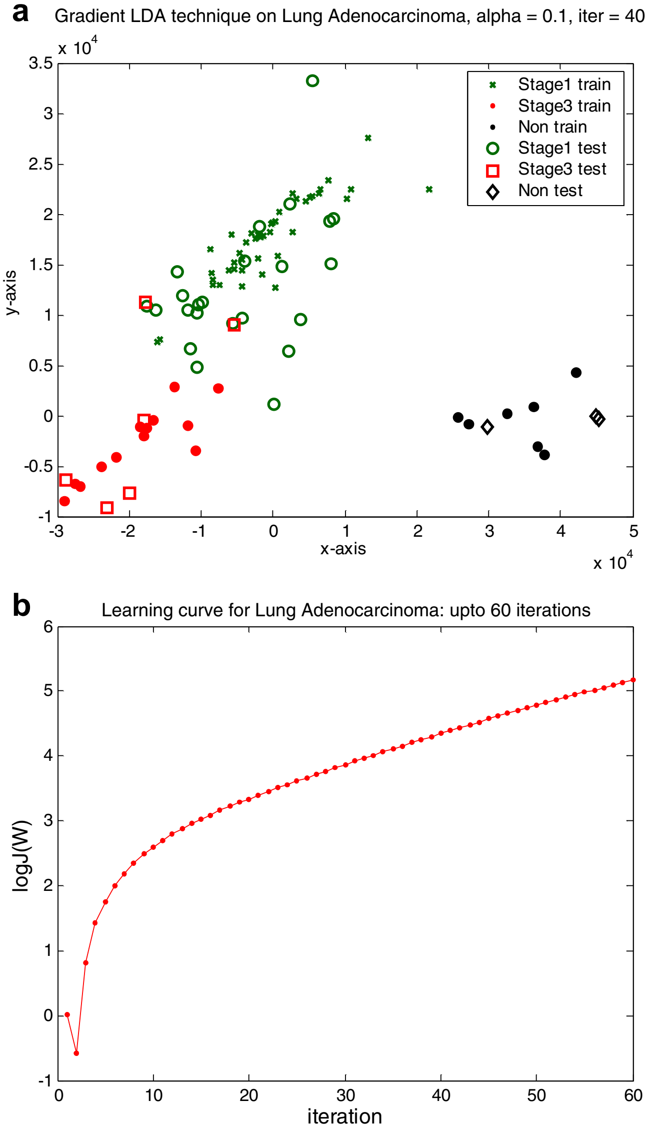

Fig. 4. (a) Classification using GLDA technique on lung adenocarcinoma dataset and (b) learning curve for lung adenocarcinoma dataset.

Table 3
Misclassification on lung adenocarcinoma dataset

Method	Number of misclassified samples
HC-TSP [24]	9
k-NN [24]	8
HC-*k*-TSP [24]	7
Decision Trees [24]	7
Naïve Bayes [24]	6
1-vs-1-SVM [24]	4
GLDA	2

3-dimensional subspace. The test samples are also shown in this figure. They are close to their corresponding train samples. Fig. 3b illustrates the learning curve on SRBCT dataset for $\alpha = 0.15$. The convergence was reached after 16 iterations. The *k*-nearest neighbour classifier is used for classifying the reduced (3-dimensional) vector into four cancer classes. The *k* value was found to be one. The GLDA technique is compared with other techniques in Table 2 and it can be observed that GLDA is giving zero misclassification error for SRBCT dataset.

The classification results using GLDA technique on lung adenocarcinoma dataset are shown in Fig. 4. The dataset contains 3 types of samples namely stage 1 tumor, stage 3 tumor and non-neoplastic lung type. The dimensionality of samples is reduced from 7129-dimensional space to 2-dimensional subspace. The projection of samples on 2-dimensional subspace is depicted in Fig. 4a. It can be seen that the training data of different classes are well separated. However, two test samples of stage 3 tumor are observed in the vicinity of train samples of stage 1 tumor type which would lead to non-zero misclassification error. Fig. 4b depicts the learning curve on lung adenocarcinoma dataset for $\alpha = 0.1$. The convergence was reached after 40 iterations. The *k*-nearest neighbour classifier is used in classifying the reduced (2-dimensional) vector into three classes. The value of *k* was found to be 7. The number of misclassified samples resulting from the GLDA technique is listed in Table 3. For comparison, we also give in this table the results from Tan et al. [24]. Note that Tan et al. [24] have randomly split the lung adenocarcinoma dataset into 64 training samples and 32 testing samples. Thus the splitting of this dataset into the training and test data used in our study may not be same as that used by Tan et al. [24], nonetheless, Table 3 gives a general idea of comparative performance of the techniques. The GLDA technique produces only two misclassification error which is lower than all the other techniques listed in Table 3.

In summary, the GLDA technique exhibits better classification performance on all of the three datasets studied in this paper than the other techniques.

5. Conclusions

In this paper, we have investigated the use of the LDA technique to reduce the dimensionality of the feature space for cancer classification using microarray gene expression data. Due to SSS problem, the conventional LDA technique cannot be applied directly on the microarray data. In order to overcome this limitation, we investigate the usefulness of the gradient LDA (GLDA) technique for reducing the dimensionality. The resulting lower-dimensional feature vector is classified using the *k*-nearest neighbour classifier. The GLDA technique is applied to three different microarray datasets. We have shown that this technique produces quite encouraging results on gene expression datasets.

References

[1] A. Antoniadis, S. Lambert-Lacroix, F. Leblanc, Effective dimension reduction methods for tumor classification using gene expression data, Bioinformatics 19 (2003) 563–570.
[2] K. Bae, B.K. Mallick, Gene selection using a two-level hierarchical Bayesian model, Bioinformatics 20 (18) (2004) 3423–3430.
[3] 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, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nature Medicine 8 (2002) 816–824. Data Source: <http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html> .
[4] N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[5] L. Breiman, Out-of-bag estimation, Technical Report, Statistics Department, University of California, Berkeley, 1996.
[6] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (2000) 1713–1726.
[7] R.O. Duda, P.E. Hart, Pattern classification and scene analysis, Wiley, New York, 1973.
[8] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Technical Report # 576, Department of Statistics, University of California, Berkeley, 2000.
[9] Y. Freund, R.E. Schapire, A decision-theoretic generation of on-line learning and application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139.
[10] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press Inc., Hartcourt Brace Jovanovich, Publishers, 1990.
[11] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.
[12] D. Geman, C. d'Avignon, D.Q. Naiman, R.L. Winslow, Classifying gene expression profiles from pairwise mRNA comparisons, Statistical Applications in Genetics and Molecular Biology 3 (1) (2004) (article 19).

[13] 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, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537. Data Source <http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43> .

[14] I. Guyon, J. Weston, S. Barnhill, Gene selection for cancer classification using Support Vector Machines, Machine Learning 46 (2002) 389–422.

[15] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network, Nature Medicine 7 (2001) 673–679. Data Source: <http://research.nhgri.nih.gov/microarray/Supplement/> .

[16] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach, Bioinformatics 19 (1) (2003) 90–97.

[17] X. Leng, H.-G. Müller, Classification using functional data analysis for temporal gene expression data, Bioinformatics 22 (1) (2006) 68–76.

[18] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (1) (2002) 39–50.

[19] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[20] A. Sharma, K.K. Paliwal, G.C. Onwubolu, Class-dependent PCA, MDC and LDA: a combined classifier for pattern classification, Pattern Recognition 39 (2006) 1215–1229.

[21] A. Sharma, K.K. Paliwal, A gradient linear discriminant analysis for small sample sized problem, Neural Processing Letters 27 (1) (2008) 17–24.

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

[23] D.L. Swets, J. Weng, Using discriminative eigenfeatures for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 831–836.

[24] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (20) (2005) 3896–3904.

[25] Y. Tan, L. Shi, W. Tong, C. Wang, Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data, Nucleic Acids Research 33 (1) (2005) 56–65.

[26] K.Y. Yeung, R.E. Bumgarner, A.E. Raftery, Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data, Bioinformatics 21 (10) (2005) 2394–2402.

[27] H.H. Zhang, J. Ahn, X. Lin, C. Park, Gene selection using support vector machine with non-convex penalty, Bioinformatics 22 (1) (2006) 88–95.

[28] X. Zhou, K.-Y. Liu, S.T.C. Wong, Cancer classification and prediction using logistic regression with Bayesian gene selection, Journal of Biomedical Informatics 37 (2004) 249–259.

Alok Sharma received the BTech degree from the University of the South Pacific (USP), Fiji in 2000, MEng degree from Griffith University, Australia with academic excellence award in 2001 and PhD degree in the area of Pattern Recognition from Griffith University in 2006. He participated in various projects carried out in conjunction with Motorola (Sydney), Auslog Pty. Ltd. (Brisbane), CRC Micro Technology (Brisbane) and French Embassy (Suva). His research interests include pattern recognition, computer security and human cancer classification. He reviewed several articles from the journals like IEEE Transactions on Neural Networks, IEEE Transaction on Systems, Man, and Cybernatics, Part A: Systems and Humans, IEEE Journal on Selected Topics in Signal Processing, IEEE Transactions on Knowledge and Data Engineering, Computers & Security and Pattern Recognition. Presently he is serving as an academic and Head of Division of the Division of Electrical/Electronics, School of Engineering and Physics, USP.

Kuldip K. Paliwal received the B.S. degree from Agra University, Agra, India, in 1969, the M.S. degree from Aligarh Muslim University, Aligarh, India, in 1971 and the Ph.D. degree from Bombay University, Bombay, India, in 1978.

He has been carrying out research in the area of speech processing since 1972. He has worked at a number of organizations including Tata Institute of Fundamental Research, Bombay, India, Norwegian Institute of Technology, Trondheim, Norway, University of Keele, UK, AT&T Bell Laboratories, Murray Hill, New Jersey, USA, AT&T Shannon Laboratories, Florham Park, New Jersey, USA, and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Microelectronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition and artificial neural networks. He has published more than 250 papers in these research areas.

Dr. Paliwal is a Fellow of Acoustical Society of India. He has served the IEEE Signal Processing Society's Neural Networks Technical Committee as a founding member from 1991 to 1995 and the Speech Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing during the periods 1994–1997 and 2003–2004. He also served as Associate Editor of the IEEE Signal Processing Letters from 1997 to 2000. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000). He has co-edited two books: "Speech Coding and Synthesis" (published by Elsevier), and "Speech and Speaker Recognition: Advanced Topics" (published by Kluwer). He has received IEEE Signal Processing Society's best (senior) paper award in 1995 for his paper on LPC quantization. He is currently serving the Speech Communication journal (published by Elsevier) as its Editor-in-Chief.