A Feature Selection Method Using Improved Regularized Linear Discriminant Analysis

Alok Sharma[1,2,3], Kuldip K. Paliwal[2], Seiya Imoto[1], Satoru Miyano[1]
[1]Laboratory of DNA Information Analysis, University of Tokyo, Japan
[2]School of Engineering, Griffith University, Australia
[3]School of Engineering and Physics, University of the South Pacific, Fiji

Abstract

Investigation of genes, using data analysis and computer based methods, has gained widespread attention in solving human cancer classification problem. DNA microarray gene expression datasets are readily utilized for this purpose. In this paper, we propose a feature selection method using improved regularized linear discriminant analysis technique to select important genes, crucial for human cancer classification problem. The experiment is conducted on several DNA microarray gene expression datasets and promising results are obtained when compared with several other existing feature selection methods.

Introduction

Feature selection methods play significant role in identifying crucial genes related to human cancers. It helps in understanding the gene regulation mechanism of cancer heterogeneity. DNA microarray gene expression data, consisting of several thousands of gene expression profiles, has been used widely in the past for cancer classification problem (Golub et al., 1999; Hastie et al., 2001; Khan et al., 2001; Armstrong, 2002). The high feature dimensionality (i.e., number of gene expression profiles) compared to the low number of samples, degrades the generalization performance of the classifier and increases its computational complexity. This problem is known as small sample size (SSS) problem (Fukunaga, 1990). These datasets along with feature selection methods provide vital information and assistance in comprehending biological and clinical characteristics. Since not all the genes are associated to cancer classification task, it is necessary to remove unimportant genes using feature selection or computational data analysis methods.

Various feature selection methods have been developed (Golub et al., 1999; Furey et al., 2000; Guyon et al., 2002; Li and Wong, 2003; Tan and Gilbert, 2003; Ding and Peng, 2003; Cong et al., 2005; Wang and Gehan, 2005; Banerjee et al., 2007; Pavlidis et al., 2001; Thomas et al., 2001; Pan, 2002; Dudoit et al., 2002; Saeys et al., 2007; Nie et al., 2010; Sharma et al., 2011, 2012a, 2012b, 2012c; Wu et al., 2011; Sharma et al., 2013a & 2013b), which can be broadly categorized into two main groups: filter methods and wrapper methods. The filter methods are classifier independent whereas the wrapper methods are classifier dependent. Filter-based methods are computationally economical and follow an open-loop approach: the selection of genes is independent of the classifier. Therefore, the relevance of the extracted genes is obtained from a scoring procedure that uses intrinsic properties of the genes' expression profiles. Wrapper-based methods (like SVM-RFE[1]) can provide high classification accuracy but are computationally intensive and follow closed-loop approaches that depend on the classifier for gene selection. Although wrapper-based methods yield high classification accuracy, the gene sets they select do not necessarily possess biologically or clinically relevant attributes.

In this paper, we propose a feature selection method using regularized linear discriminant analysis (RLDA) technique (Friedman, 1989). This feature selection method falls under the filter method category as it does not require a classifier during training process to select features.

RLDA technique is one of the few pioneering techniques in the pattern classification literature. RLDA technique is used in the cases where SSS exist. In RLDA, a small perturbation, known as the regularization parameter α, is added to within-class scatter matrix $\mathbf{S}_W$, to overcome SSS problem. The matrix $\mathbf{S}_W$ is approximated by $\mathbf{S}_W + \alpha\mathbf{I}$ and the orientation matrix is computed by eigenvalue decomposition (EVD) of $(\mathbf{S}_W + \alpha\mathbf{I})^{-1}\mathbf{S}_B$, where $\mathbf{S}_B$ is between-class scatter matrix. RLDA has been applied in face recognition and bioinformatics area (Dai and Yuen, 2003, 2007; Guo et al., 2007). In RLDA, it can be computationally expensive to find the optimum value of the parameter α as heuristic approach (e.g. cross-validation procedure, Hastie et al., 2001) is applied. The value of the parameter could be sensitive and noisy especially when the number of training samples is scarce. In human cancer classification problem, the DNA microarray gene expression datasets, usually have very limited number of training samples which could adversely affect the classification performance of the RLDA technique.

In order to find the gene subset associated with human cancers, we first determine the value of α for RLDA technique without using any heuristic approach. We call our procedure as improved RLDA technique. We use improved RLDA technique recursively to obtain crucial genes important for cancer classification task. The proposed feature

[1] SVM-RFE (Guyon et al., 2002) is a wrapper based method. It is an iterative method which works backward from an initial set of features. The SVM aims to find maximum margin hyperplane between the two classes to minimize classification error using some kernel function.

selection method has been applied on several DNA microarray gene expression datasets and promising results have been obtained.

In the past, SVM has also applied recursively in SVM-RFE method (Guyon et al., 2002) to select features. SVM-RFE is a wrapper based method. It is an iterative method which works backward from an initial set of features. The SVM aims to find maximum margin hyperplane between the two classes to minimize classification error using some kernel function. The selection of features by SVM-RFE is computationally intensive. It has some other drawbacks as well due to applying maximum margin criterion between two classes (Zhou et al., 2010). On the other hand, RLDA based recursive feature selection method, separates the two classes by 1) shrinking within class variance, and 2) increasing the between class variance.

Basic descriptions

In this section we describe the basic notations used in the paper. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ denote n training samples (or feature vectors) in a d-dimensional space having class labels $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, where $\omega \in \{1,2,\dots,c\}$ and c is the number of classes. The dataset $\mathbf{X}$ can be subdivided into c subsets $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_c$, where $\mathbf{X}_j$ belongs to class j and consists of n_j number of samples such that $n = \sum_{j=1}^{c} n_j$. The data subset $\mathbf{X}_j \subset \mathbf{X}$ and $\mathbf{X}_1 \cup \mathbf{X}_2 \cup \ldots \cup \mathbf{X}_c = \mathbf{X}$. If $\boldsymbol{\mu}_j = 1/n_j \sum_{\mathbf{x}\in\mathbf{X}_j} \mathbf{x}$ is the centroid of $\mathbf{X}_j$ and $\boldsymbol{\mu} = 1/\mathrm{n} \sum_{\mathbf{x}\in\mathbf{X}} \mathbf{x}$ is the centroid of $\mathbf{X}$, then the total scatter matrix $\mathbf{S}_T$, within-class scatter matrix $\mathbf{S}_W$ and between-class scatter matrix $\mathbf{S}_B$ are defined as (Duda and Hart, 1973; Sharma and Paliwal, 2008a, 2008b; Xu and Yan, 2009; Sharma and Paliwal, 2012; Huang, 2012a, 2012b)

$$\mathbf{S}_T = \textstyle\sum_{\mathbf{x}\in\mathbf{X}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}},$$

$$\mathbf{S}_W = \textstyle\sum_{j=1}^{c} \sum_{\mathbf{x}\in\mathbf{X}_j} (\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}},$$

and $$\mathbf{S}_B = \textstyle\sum_{j=1}^{c} n_j\, (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}}.$$

In SSS problem, $d > n$, which will make scatter matrices singular. Let r_t be the rank of $\mathbf{S}_T$ matrix. The eigenvector decomposition of $\mathbf{S}_T$ can be given as

$$\mathbf{S}_T = [\mathbf{U}_1, \mathbf{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_T & \\ & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^{\mathrm{T}} \\ \mathbf{U}_2^{\mathrm{T}} \end{bmatrix}, \qquad (1)$$

where $\mathbf{U}_1 \in \mathbb{R}^{d\times r_t}$ corresponds to eigenvalues $\boldsymbol{\Lambda}_{\mathrm{T}}$ and $\mathbf{U}_2 \in \mathbb{R}^{d\times(d-r_t)}$ corresponds to the zero eigenvalues. The matrix $\mathbf{U}_1$ is the range space of $\mathbf{S}_T$ and the matrix $\mathbf{U}_2$ is the null space of $\mathbf{S}_T$. Since the null space of $\mathbf{S}_T$ does not contain any discriminant information (Huang et al., 2002), the dimensionality can be reduced from

d-dimensional space to r_t-dimensional space by applying principal component analysis (PCA) (Fukunaga, 1990; Sharma and Paliwal, 2007) as a pre-processing step. The range space of $\mathbf{S}_T$ matrix, $\mathbf{U}_1 \in \mathbb{R}^{d \times r_t}$, will be used as a transformation matrix. In the reduced dimensional space the scatter matrices can be computed by: $\mathbf{S}_W \leftarrow \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_W \mathbf{U}_1$ and $\mathbf{S}_B \leftarrow \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_B \mathbf{U}_1$. After this procedure $\mathbf{S}_W \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{S}_B \in \mathbb{R}^{r_t \times r_t}$ are reduced dimensional within-class scatter matrix and reduced dimensional between-class scatter matrix, respectively.

Improved RLDA technique for feature selection

In RLDA, the regularization of within-class scatter matrix $\mathbf{S}_W$ is carried out by adding a perturbation term α to the diagonal elements of $\mathbf{S}_W$; i.e., $\hat{\mathbf{S}}_W = \mathbf{S}_W + \alpha \mathbf{I}$. The addition of α will make within-class scatter non-singular and invertible. This would help to maximize the modified Fisher's criterion

$$J(\mathbf{w}, \alpha) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_B \mathbf{w}}{\mathbf{w}^{\mathrm{T}} (\mathbf{S}_W + \alpha \mathbf{I}) \mathbf{w}}, \qquad (2)$$

where $\mathbf{w} \in \mathbb{R}^{r_t \times 1}$ is the orientation vector. In order to avoid any heuristic approach in the determination of the parameter α, we solve equation 2 in the following manner. Let us denote function $f = \mathbf{w}^{\mathrm{T}} \mathbf{S}_B \mathbf{w}$ and a constraint function $g = \mathbf{w}^{\mathrm{T}} (\mathbf{S}_W + \alpha \mathbf{I}) \mathbf{w} - c = 0$, where $c > 0$ be any constant. To find the constrained relative-maximum of function f under constrained curve g, we can use the method of Lagrange multipliers (Anton, 1995) as follows:

$$\frac{\partial f}{\partial \mathbf{w}} = \lambda \frac{\partial g}{\partial \mathbf{w}}, \qquad (3)$$

where $\lambda \neq 0$ is the Lagrange's multiplier. Equation 3 is the Lagrange's function where we are interested in finding the parameters $(\mathbf{w}, \lambda)$ that maximizes function f under the constrained curve g. Substituting $f = \mathbf{w}^{\mathrm{T}} \mathbf{S}_B \mathbf{w}$ and $g = \mathbf{w}^{\mathrm{T}} (\mathbf{S}_W + \alpha \mathbf{I}) \mathbf{w} - c$ in equation 3, we get

$$2\mathbf{S}_B \mathbf{w} = \lambda (2\mathbf{S}_W \mathbf{w} + 2\alpha \mathbf{w}),$$

or
$$\left(\tfrac{1}{\lambda} \mathbf{S}_B - \mathbf{S}_W\right) \mathbf{w} = \alpha \mathbf{w}. \qquad (4)$$

The value of $\alpha \mathbf{w}$ can be substituted in the constraint function g, this will give us,

$$\mathbf{w}^{\mathrm{T}} \mathbf{S}_B \mathbf{w} = \lambda c. \qquad (5)$$

Also from the constraint function $\mathbf{w}^{\mathrm{T}} (\mathbf{S}_W + \alpha \mathbf{I}) \mathbf{w} - c = 0$, we get $\mathbf{w}^{\mathrm{T}} \hat{\mathbf{S}}_W \mathbf{w} = c$. Dividing this term in equation 5, we get

$$\lambda = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{B}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \hat{\mathbf{S}}_W \mathbf{w}}. \qquad (6)$$

We can observe the following things from equation 6: 1) the left-hand term is the

Lagrange's multiplier (in equation 4), and 2) the right-hand side is same as the Fisher's modified criterion defined in equation 2. In order to obtain the value of λ in equation 6, we need to estimate $\hat{\mathbf{S}}_W$. If the matrix is not regularize (i.e., $\alpha = 0$) then $\hat{\mathbf{S}}_W = \mathbf{S}_W$. By this substitution, we can obtain approximate value of λ by maximizing $\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}/\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$. Now to find the maximum value of $\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}/\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$, we must have eigenvector $\mathbf{w}$ corresponding to the leading eigenvalue of $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$. However, since $\mathbf{S}_W$ is singular and non-invertible, $\mathbf{S}_W^{+}$ can be used in place of $\mathbf{S}_W^{-1}$, where $\mathbf{S}_W^{+}$ is the pseudoinverse of $\mathbf{S}_W$. From the EVD of $\mathbf{S}_W^{+}\mathbf{S}_B$, we can find λ_{max} which is the largest eigenvalue of $\mathbf{S}_W^{+}\mathbf{S}_B$. The value of λ_{max} can be substituted in equation 4 (where $\lambda = \lambda_{max}$), this will enable us to find the value of α by doing EVD of $(\frac{1}{\lambda}\mathbf{S}_B - \mathbf{S}_W)$. If $r_b = rank(\mathbf{S}_B)$ then EVD of $(\frac{1}{\lambda}\mathbf{S}_B - \mathbf{S}_W)$ will give r_b finite eigenvalues. Since the leading eigenvalue will correspond to the most discriminant eigenvector (Fukunaga, 1990; Sharma and Paliwal, 2007), α is taken to be the leading eigenvalue. Once the value of α is determined, the orientation vector $\mathbf{w}$ can be solved from

$$(\mathbf{S}_W + \alpha\mathbf{I})^{-1}\mathbf{S}_B\mathbf{w} = \gamma\mathbf{w}. \qquad (7)$$

It can be shown from Lemma 1 that for improved RLDA technique, its maximum eigenvalue is approximately equal to the highest (finite) eigenvalue of Fisher's criterion.

Lemma 1: *The highest eigenvalue of improved RLDA is approximately equivalent to the highest (finite) eigenvalue of Fisher's criterion.*

Proof 1: From equation 7,

$$\mathbf{S}_B\mathbf{w}_j = \gamma_j(\mathbf{S}_W + \alpha\mathbf{I})\mathbf{w}_j, \qquad (8)$$

where α is the maximum eigenvalue of $(1/\lambda_{max}\mathbf{S}_B - \mathbf{S}_W)$ (from equation 4); $\lambda_{max} \geq 0$ is approximately the highest eigenvalue of Fisher's criterion $\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}/\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$ (since λ_{max} is the largest eigenvalue of $\mathbf{S}_W^{+}\mathbf{S}_B$) (Liu et al., 2007); $j = 1 \ldots r_b$ and $r_b = rank(\mathbf{S}_B)$. Substituting $\alpha\mathbf{w} = (1/\lambda_{max}\mathbf{S}_B - \mathbf{S}_W)\mathbf{w}$ (from equation 4, where $\lambda = \lambda_{max}$) into equation 8, we get,

$$\mathbf{S}_B\mathbf{w}_m = \gamma_m\mathbf{S}_W\mathbf{w}_m + \gamma_m(1/\lambda_{max}\mathbf{S}_B - \mathbf{S}_W)\mathbf{w}_m,$$

or $\quad (\lambda_{max} - \gamma_m)\mathbf{S}_B\mathbf{w}_m = 0.$

where $\gamma_m = \max(\gamma_j)$ and $\mathbf{w}_m$ is the corresponding eigenvector. Since $\mathbf{S}_B\mathbf{w}_m \neq 0$ (from equation 5), $\gamma_m = \lambda_{max}$ and $\gamma_j < \lambda_{max}$, where $j \neq m$. This concludes the proof.

Corollary 1: The value of regularization parameter is non-negative; i.e., $\alpha \geq 0$ for $r_w \leq r_t$, where $r_t = rank(\mathbf{S}_T)$ and $r_w = rank(\mathbf{S}_W)$.

Proof. Please see Appendix-III.

Computing equation 7 for all the values of γ will give the orientation matrix $\mathbf{W} \in \mathbb{R}^{r_t \times r_b}$, having $\mathbf{w}$ as its column vectors. The orientation matrix $\mathbf{W}$ is in r_t-dimensional space, however, it can be transformed to d-dimensional space by $\mathbf{W} \leftarrow \mathbf{U}_1\mathbf{W}$. Therefore, we get $\mathbf{W} \in \mathbb{R}^{d \times r_b}$. Let a column vector $\mathbf{w} \in \mathbf{W}$ be used to transform d-dimensional space to 1-dimensional space and $\mathbf{x} \in \mathbf{X}$ be any feature vector, we have

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x},$$

or
$$y = \textstyle\sum_{i=1}^{d} w_i x_i, \qquad (9)$$

where w_i and x_i are the elements of $\mathbf{w}$ and $\mathbf{x}$, respectively. It can be envisaged that if $|w_i x_i| \approx 0$ (where $|\cdot|$ is the absolute value), then the ith element is not contributing for the value of y in equation 9; i.e., it can be discarded without sacrificing much information. This concept can be extended for the orientation matrix $\mathbf{W}$ and dataset $\mathbf{X}$ as

$$z_i = \textstyle\sum_{k=1}^{r_b} \sum_{j=1}^{n} |w_{ik} x_{ij}| \qquad (10)$$

where $i = 1,2,\ldots,d$. If $z_i \approx 0$, then ith feature can be discarded. Equation 10 can be applied recursively to discard unimportant features as follows:

Step 0. Define $q \in (n, d)$[2] and set $l = d$.
Step 1. Compute $\mathbf{W} \in \mathbb{R}^{l \times r_b}$ (see Table 1).
Step 2. Compute z_i using equation 10 for $i = 1,2,\ldots,l$.
Step 3. Sort z_i in descending order; i.e., if $s = sort(z_i)$ then $s_1 > s_2 > \cdots > s_l$.
Step 4. Discard least important feature corresponding to s_l. Let the cardinality of the remaining feature set be $l-1$ and data subset be $\mathbf{X}_{l-1} \in \mathbb{R}^{l \times n}$.
Step 5. Conduct $\mathbf{X} \leftarrow \mathbf{X}_{l-1}$ and $l \leftarrow l-1$.
Step 6. Continue Steps 1-5 until $l = q$.

The above process will give q-features with the data subset $\mathbf{X}_q \in \mathbb{R}^{q \times n}$, which can be used by a classifier to obtain classification performance.

Table 1: Computation of the orientation matrix $\mathbf{W}$ using improved RLDA technique.

[2] Since RLDA or Improved RLDA is a method for solving small sample size (SSS) problem, the value of q has to be in (n, d).

Step 1. Compute range space of total scatter matrix $\mathbf{S}_T$, $\mathbf{U}_1 \in \mathbb{R}^{d \times r_t}$, by applying PCA, where $r_t = rank(\mathbf{S}_T)$. Using $\mathbf{U}_1$, compute between-class scatter matrix and within-class scatter matrix in r_t dimensional space: $\mathbf{S}_B \leftarrow \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_B \mathbf{U}_1$ and $\mathbf{S}_W \leftarrow \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_W \mathbf{U}_1$, where $\mathbf{S}_B \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{S}_W \in \mathbb{R}^{r_t \times r_t}$.

Step 2. Perform EVD of $\mathbf{S}_W^{+} \mathbf{S}_B$ to find the highest eigenvalue λ_{max}.

Step 3. Perform EVD of $(1/\lambda_{max} \mathbf{S}_B - \mathbf{S}_W)$ to find its highest eigenvalue, denote it as α.

Step 4. Perform EVD of $(\mathbf{S}_W + \alpha \mathbf{I})^{-1} \mathbf{S}_B$ to find r_b eigenvectors $\mathbf{w}_j \in \mathbb{R}^{r_t \times 1}$ corresponding to the leading eigenvalues, where $r_b = rank(\mathbf{S}_B)$. The eigenvectors $\mathbf{w}_j$ are column vectors of the orientation matrix $\mathbf{W}' \in \mathbb{R}^{r_t \times r_b}$.

Step 5. Compute the orientation matrix $\mathbf{W} \in \mathbb{R}^{d \times r_b}$ in d-dimensional space: $\mathbf{W} = \mathbf{U}_1 \mathbf{W}'$.

The computational requirement for Step 1 of the technique (Table 1) would be $O(dn^2)$; for Step 2 would be $O(n^3)$; for Step 3 would be $O(n^3)$; for Step 4 would be $O(n^3)$; and, for Step 5 would be $O(dn^2)$. Therefore, the total estimated for SSS case $(d \gg n)$ would be $O(dn^2)$. If the q features are to be selected from the total d features then total estimated computational complexity would be $O(dn^2(d-l))$.

Experimentation

In this experiment we have utilized three DNA microarray gene expression datasets[3]. The description of these datasets is given as follows:

SRBCT dataset (Khan et al., 2001): 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 test set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively.

MLL Leukemia dataset (Armstrong et al., 2002): this dataset has 3 classes namely ALL,

[3] Most of the datasets are downloaded from the Kent Ridge Bio-medical Dataset (KRBD) (http://datam.i2r.a-star.edu.sg/datasets/krbd/). The datasets are transformed or reformatted and made available by KRBD repository and we have used them without any further preprocessing. Some datasets which are not available on KRBD repository are downloaded and directly used from respective authors' supplement link. The URL addresses for all the datasets are given in the Reference Section.

MLL and AML. The training set contains 57 leukemia samples (20 ALL, 17 MLL and 20 AML) whereas the test set contains 15 samples (4 ALL, 3 MLL and 8 AML). The dimension of the MLL dataset is 12582.

Acute Leukemia dataset (Golub et al., 1999): this dataset consists of DNA microarray gene expression data of human acute leukemia for cancer classification. Two types of acute leukemia 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 test 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.

The classification performance of the proposed feature selection method has been gauged by using the above three datasets. Tables 2, 3 and 4 show classification accuracy of the proposed method compared with several other existing feature selection methods on the SRBCT, MLL and Acute Leukemia datasets, respectively[4]. Four classifiers from WEKA (http://www.cs.waikato.ac.nz/ml/weka/) used are J4.8, Naïve Bayes, kNN (where $k = 1$) and SVM pairwise. The classification accuracy for the SRBCT and MLL datasets is obtained from Tao et al. 2004. For all the datasets, the features are ranked by Rankgene program (Su et al., 2003). The Rankgene program computes the features for the following feature selection methods: Information gain, Twoing rule, Sum minority, Max minority, Gini index, Sum of variances, t-statistic and One-dimensional SVM (Su et al., 2003). For all the datasets 150 genes are selected as selected by Tao et al., 2004. In addition, Lasso (Tibshirani, 1996) and filter MRMR (Peng et al., 2005) are used for feature selection. The Lasso method deflates the collinearity effect on the features. It produces sparse parameters that can be used to identify important genes. The number of features selected by Lasso on SRBCT, MLL and Acute Leukamia is 38, 39 and 16[5], respectively. The filter MRMR method select features based on maximal statistical dependency criterion based on mutual information. It can be observed from Table 2 that

[4] The cross-validation based results are shown in Appendix-I Section. The comparison of improved RLDA with different values of regularization parameter has been shown in Appendix-II Section.

[5] Note that for all the feature selection methods except Lasso method the number of selected features is 150 (in Tables 2,3 and 4). The Lasso method itself obtains the optimal number of selected features and therefore cannot be adjusted for a predefined number of selected features.

the proposed method achieves 75% classification accuracy using the J4.8 classifier; 90% classification accuracy using the Naïve Bayes classifier; 95% classification accuracy using the kNN classifier and 100% classification accuracy by the SVM pairwise classifier. In the three out of four cases the classification accuracy obtained by improved RLDA is the highest. Similarly, the classification accuracy on the MLL dataset (Table 3) is the highest for improved RLDA in three out of four cases method when compared with several other feature selection methods using four distinct classifiers. On the Acute Leukemia dataset (Table 4), the classification accuracy of improved RLDA is the highest for the J4.8 classifier (94%) and the SVM pairwise classifier (100%). In total of 12 cases (Tables 2-4), improved RLDA is giving highest results in eight cases. It can, therefore, be concluded that the proposed method is exhibiting promising results.

Table 2: The classification accuracy of various feature selection methods using four distinct classifiers on the SRBCT dataset.

	J4.8	Naïve Bayes	kNN	SVM pairwise
Baseline accuracy	37%	37%	37%	37%
Information gain	68%	68%	90%	90%
Twoing rule	64%	73%	86%	82%
Sum minority	68%	68%	90%	86%
Max minority	46%	78%	90%	90%
Gini index	64%	78%	90%	90%
Sum of variances	54%	64%	90%	86%
t-statistic	54%	64%	90%	86%
One dimensional SVM	54%	64%	90%	86%
Lasso	90%	70%	80%	75%
Filter MRMR	65%	35%	55%	85%
Improved RLDA	75%	90%	95%	100%

Table 3: The classification accuracy of various feature selection methods using four distinct classifiers on the MLL dataset.

	J4.8	Naïve Bayes	kNN	SVM pairwise
Baseline accuracy	35%	35%	35%	35%
Information gain	60%	74%	86%	100%
Twoing rule	60%	86%	86%	100%
Sum minority	68%	26%	80%	80%
Max minority	74%	34%	74%	80%
Gini index	60%	68%	86%	100%
Sum of variances	60%	54%	86%	100%
t-statistic	60%	54%	86%	100%
One dimensional SVM	60%	54%	86%	100%
Lasso	87%	100%	93%	93%
Filter MRMR	100%	100%	93%	100%
Improved RLDA	100%	93%	100%	100%

Table 4: The classification accuracy of various feature selection methods using four distinct classifiers on the Acute Leukemia dataset.

	J4.8	Naïve Bayes	kNN	SVM pairwise
Baseline accuracy	71%	71%	71%	71%
Information gain	91%	100%	97%	97%
Twoing rule	91%	97%	97%	97%
Sum minority	91%	97%	97%	97%
Max minority	91%	97%	97%	97%
Gini index	91%	97%	97%	97%
Sum of variances	91%	97%	97%	97%
t-statistic	91%	100%	97%	97%
One dimensional SVM	91%	85%	88%	97%
Lasso	91%	94%	85%	91%
Filter MRMR	65%	71%	74%	86%
Improved RLDA	94%	94%	85%	100%

Next, we considered different number of selected features by Improved RLDA and several feature selection method, and shown the evolution of the performance of the classifiers with respect to the number of selected features. The results are shown in Tables 5, 6 and 7. It can be observed from the Tables 5-7 that in most of the cases the average classification accuracy for Improved RLDA is consistently higher than other feature selection methods.

Table 5: The classification accuracy as a function of the number of selected features of Improved RLDA and several feature selection methods using four distinct classifiers on the SRBCT dataset.

Feature selection + classifier	10% of features	20% of features	30% of features	Average classification accuracy
Information gain + J4.8	65%	65%	65%	81.7%
Information gain + Naïve Bayes	85%	65%	55%	
Information gain + kNN	100%	90%	90%	
Information gain + SVM	100%	100%	100%	
Twoing rule + J4.8	65%	65%	65%	82.1%
Twoing rule + Naïve Bayes	85%	70%	55%	
Twoing rule + kNN	100%	90%	90%	
Twoing rule + SVM	100%	100%	100%	
Sum minority + J4.8	60%	65%	65%	79.6%
Sum minority + Naïve Bayes	75%	55%	55%	
Sum minority + kNN	100%	95%	85%	
Sum minority + SVM	100%	100%	100%	
Max minority + J4.8	65%	65%	65%	83.3%
Max minority + Naïve Bayes	95%	65%	65%	
Max minority + kNN	100%	90%	90%	
Max minority + SVM	100%	100%	100%	
Gini index + J4.8	65%	75%	75%	85.8%
Gini index + Naïve Bayes	90%	70%	65%	
Gini index + kNN	100%	95%	95%	
Gini index + SVM	100%	100%	100%	
Sum of variances + J4.8	65%	65%	65%	79.2%
Sum of variances + Naïve Bayes	60%	60%	55%	
Sum of variances + kNN	100%	90%	90%	
Sum of variances + SVM	100%	100%	100%	
Improved RLDA + J4.8	75%	75%	75%	88.3%
Improved RLDA + Naïve Bayes	90%	90%	70%	
Improved RLDA + kNN	95%	95%	95%	
Improved RLDA + SVM pairwise	100%	100%	100%	

Table 6: The classification accuracy as a function of the number of selected features of Improved RLDA and several feature selection methods using four distinct classifiers on the MLL dataset.

Feature selection + classifier	10% of features	20% of features	30% of features	Average classification accuracy
Information gain + J4.8	67%	67%	67%	88.5%
Information gain + Naïve Bayes	100%	100%	100%	
Information gain + kNN	87%	87%	87%	
Information gain + SVM	100%	100%	100%	
Twoing rule + J4.8	67%	67%	67%	88.5%
Twoing rule + Naïve Bayes	100%	100%	100%	
Twoing rule + kNN	87%	87%	87%	
Twoing rule + SVM	100%	100%	100%	
Sum minority + J4.8	67%	67%	67%	88.5%
Sum minority + Naïve Bayes	100%	100%	100%	
Sum minority + kNN	87%	87%	87%	
Sum minority + SVM	100%	100%	100%	
Max minority + J4.8	67%	67%	67%	88.5%
Max minority + Naïve Bayes	100%	100%	100%	
Max minority + kNN	87%	87%	87%	
Max minority + SVM	100%	100%	100%	
Gini index + J4.8	67%	67%	67%	88.5%
Gini index + Naïve Bayes	100%	100%	100%	
Gini index + kNN	87%	87%	87%	
Gini index + SVM	100%	100%	100%	
Sum of variances + J4.8	67%	67%	67%	88.5%
Sum of variances + Naïve Bayes	100%	100%	100%	
Sum of variances + kNN	87%	87%	87%	
Sum of variances + SVM	100%	100%	100%	
Improved RLDA + J4.8	100%	100%	100%	96.2%
Improved RLDA + Naïve Bayes	100%	100%	100%	
Improved RLDA + kNN	87%	87%	80%	
Improved RLDA + SVM pairwise	100%	100%	100%	

Table 7: The classification accuracy as a function of the number of selected features of Improved RLDA and several feature selection methods using four distinct classifiers on the Acute Leukemia dataset.

Feature selection + classifier	10% of features	20% of features	30% of features	Average classification accuracy
Information gain + J4.8	91%	91%	91%	90.6%
Information gain + Naïve Bayes	97%	100%	100%	
Information gain + kNN	77%	79%	79%	
Information gain + SVM	97%	94%	91%	
Twoing rule + J4.8	91%	91%	91%	89.1%
Twoing rule + Naïve Bayes	94%	97%	97%	
Twoing rule + kNN	77%	76%	79%	
Twoing rule + SVM	97%	91%	88%	
Sum minority + J4.8	91%	91%	91%	88.3%
Sum minority + Naïve Bayes	94%	97%	97%	
Sum minority + kNN	77%	73%	73%	
Sum minority + SVM	97%	91%	88%	
Max minority + J4.8	91%	91%	91%	89.2%
Max minority + Naïve Bayes	94%	97%	97%	
Max minority + kNN	77%	77%	79%	
Max minority + SVM	97%	91%	88%	
Gini index + J4.8	91%	91%	91%	88.0%
Gini index + Naïve Bayes	94%	97%	97%	
Gini index + kNN	79%	70%	70%	
Gini index + SVM	97%	91%	88%	

Sum of variances + J4.8	91%	91%	91%	89.2%
Sum of variances + Naïve Bayes	94%	97%	97%	
Sum of variances + kNN	77%	77%	79%	
Sum of variances + SVM	97%	91%	88%	
Improved RLDA + J4.8	91%	91%	91%	92.5%
Improved RLDA + Naïve Bayes	97%	100%	100%	
Improved RLDA + kNN	88%	79%	82%	
Improved RLDA + SVM pairwise	97%	97%	97%	

Furthermore, we conducted experiments to see the biological significance of the selected features by the proposed method. We use SRBCT data as a prototype to show the biological significance using Ingenuity Pathway Analysis[6]. The selected 150 features from the proposed algorithm are used for this purpose. Out of 150 genes, 10 genes were found unmapped in IPA. The top five high level biological functions obtained are shown in Figure 1. In the figure, the y-axis denotes the negative of logarithm of p-values and x-axis denotes the high level functions. Since the cancer function is of paramount interest, we investigated them further. There are 61 cancer sub-functions obtained from the experiment. Top 25 cancer sub-functions with significant p-values are shown in Table 8. In IPA, the p-value reflects the enrichment of a given function to a set of focused genes. The smaller the p-value is, the less likely that the association is random, and the more significant the association. In general p-values less than 0.05 indicate a statistically significant, non-random association. The p-value is calculated using the right-tailed Fisher exact test (IPA, Available at: http://www.ingenuity.com) (Sharma et al., 2012a; 2012b). In the table, the p-values and the number of selected genes are depicted corresponding to the selected functions. The selected genes by the proposed method provide significant p-values above the threshold (as specified in IPA). This shows that the features selected by the proposed method contain useful information for discriminatory purpose and have biological significance.

[6] Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com) is a software that helps researchers to model, analyze, and understand the complex biological and chemical systems at the core of life science research. IPA has been broadly adopted by the life science research community. IPA helps to understand complex 'omics data at multiple levels by integrating data from a variety of experimental platforms and providing insight into the molecular and chemical interactions, cellular phenotypes, and disease processes of the system. IPA provides insight into the causes of observed gene expression changes and into the predicted downstream biological effects of those changes. Even if the experimental data is not available, IPA can be used to intelligently search the Ingenuity Knowledge Base for information on genes, proteins, chemicals, drugs, and molecular relationships to build biological models or to get up to speed in a relevant area of research. IPA provides the right biological context to facilitate informed decision-making, advance research project design, and generate new testable hypotheses.

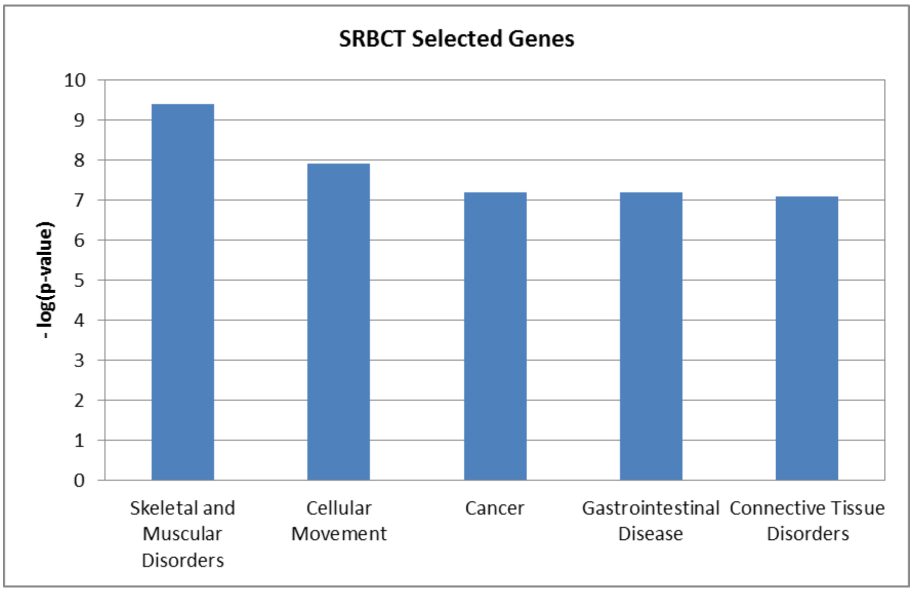

Figure 1: Top five high level biological function on selected 150 genes of SRBCT by Improved RLDA based feature selection method.

Table 8: Cancer sub-functions

Functions	p-value	# Selected Genes
metastatic colorectal cancer	6.99E-08	12
tumorigenesis	1.01E-07	62
neoplasia	5.05E-07	59
cancer	6.97E-07	58
uterine cancer	2.87E-06	19
benign tumor	3.75E-06	17
leiomyomatosis	1.06E-05	12
carcinoma	1.11E-05	47
adenocarcinoma	1.81E-05	17
gastrointestinal tract cancer	2.60E-05	24
colorectal cancer	3.46E-05	22
uterine leiomyoma	5.62E-05	10
metastasis	6.11E-05	13
genital tumor	6.69E-05	22
prostate cancer	1.42E-04	16
trisomy 8 myelodysplastic syndrome	2.25E-04	2
central nervous system tumor	2.87E-04	10
digestive organ tumor	3.21E-04	27
breast cancer	3.41E-04	20
brain cancer	4.28E-04	9
leukemia	6.88E-04	11
hematologic cancer	7.14E-04	14
endometrial carcinoma	8.86E-04	8
neuroblastoma	1.25E-03	5
hematological neoplasia	1.38E-03	15

endocrine gland tumor	1.42E-03	11
tumorigenesis of carcinoma	1.54E-03	2
B-cell leukemia	1.68E-03	6
entrance of tumor cell lines	2.04E-03	2
endometrial cancer	2.12E-03	7

We have also carried out sensitivity analysis to check the robustness of the proposed method. For this purpose, we use the SRBCT dataset as a prototype and select top 100 genes. After this selection, we contaminate the dataset by adding Gaussian noise, then applied the method again to find the top 100 genes. The generated noise levels are 1%, 2% and 5% of the standard deviation of the original gene expression values. The number of genes which are common after contamination and before contamination is noted. This contamination of data and selection of genes are repeated 20 times. The average number of genes over 20 iterations is depicted in Figure 2. It can be observed from the figure that the proposed method is able to capture the majority of the original genes in the noisy environmental condition.

In order to check the sensitivity analysis with respect to the classification accuracy, we contaminated the dataset by adding Gaussian noise (as above) and selected 150 features using the improved RLDA technique. The classification accuracy is obtained by using the SVM-pairwise classifier. The results are shown in Table 9. It can observed from Table 9 that for low level noise the degradation in classification performance is not enough. But when the noise level increases the classification accuracy deteriorates (especially on the MLL dataset and the Acute Leukemia dataset).

Table 9: Sensitivity analysis with respect to classification accuracy on the SRBCT, MLL and Acute Leukemia dataset

Noise level	SRBCT	MLL	Acute Leukemia
Without noise	100%	100%	97%
1%	100%	100%	97%
2%	100%	93%	96%
5%	100%	79%	93%
10%	100%	45%	59%

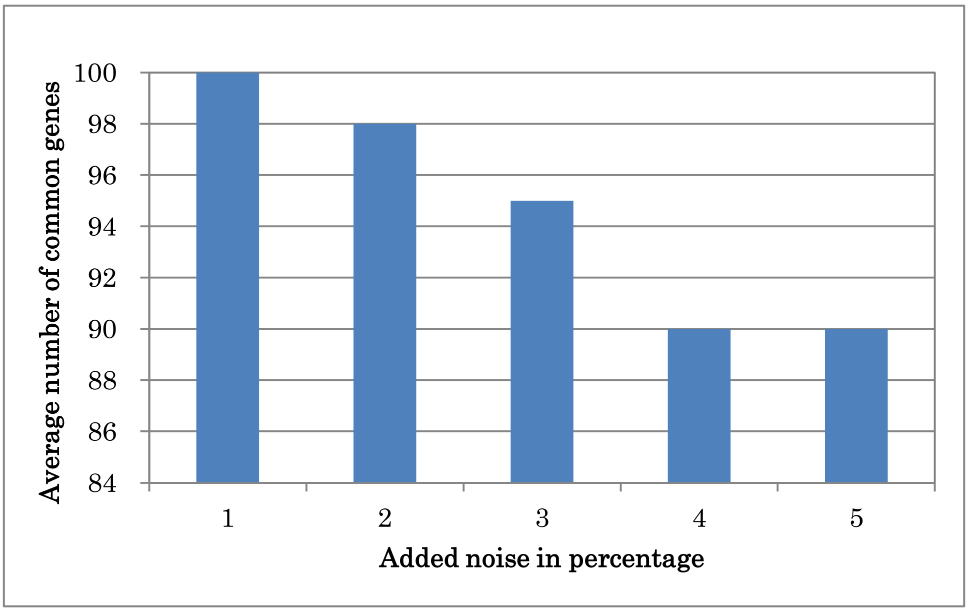

Figure 2: Sensitivity analysis for the proposed feature selection method on the SRBCT dataset at different noise levels. The y-axis depicts the average number of common genes over 20 iterations and x-axis depicts the added noise in percentage.

Next, we carried out experimentation to obtain ROC curve and AUC analysis. For the ROC curve, we use sensitivity and specificity as the two measures. The sensitivity is given as $True\ Positive/(True\ Positive + False\ Negative)$ and the specificity is given as $True\ negative/(True\ Negative + False\ Positive)$. We varied the noise level and select 150 genes using improved RLDA and then use SVM-pairwise to compute sensitivity and specificity. The ROC curve is shown in Figure 3. This curve shows the trade-off between sensitivity and specificity. The AUC provides the overall accuracy and is a useful parameter for comparing the performance. The high value of AUC parameter indicates high accuracy. The value of AUC is computed to be 0.9674 which is promising.

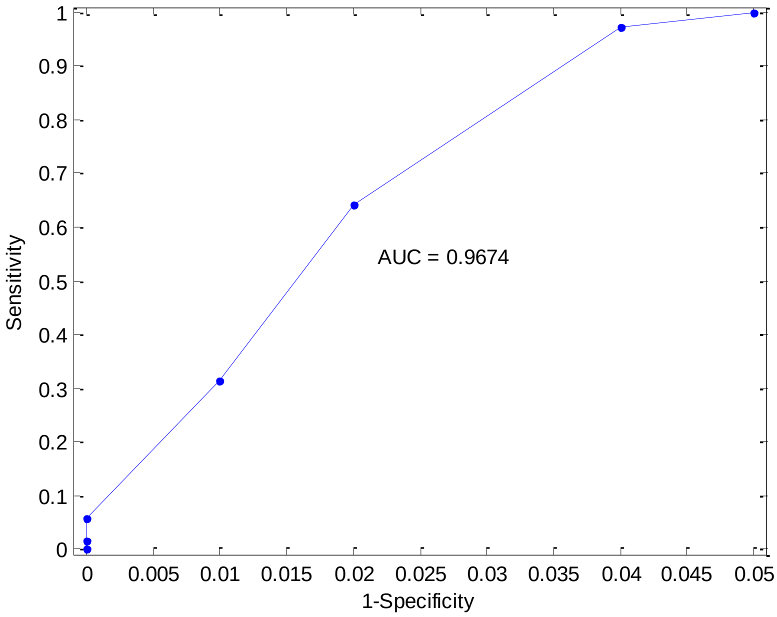

Figure 3: The ROC curve

Conclusion

In this paper, we presented a feature selection method using improved regularized linear discriminant analysis technique. Three DNA microarray gene expression datasets have been utilized to see the performance of the proposed method. It was observed that the method is achieving encouraging classification accuracy using small number of selected gene. The biological significance has also been demonstrated by performing functional analysis. Moreover, robustness of the method was exhibited by conducting sensitivity analysis and encouraging results are obtained. The sensitivity analysis with respect to classification accuracy and ROC curve have also been discussed.

Appendix I

In this section, we use cross-validation procedure to compute average classification accuracy using four distinct classifiers and the proposed feature selection method. Three datasets have been used for this purpose are SRBCT, MLL and Acute Leukemia. The classification accuracy using fold $k = 5$ and fold $k = 10$ are given in Tables A1, A2 and A3. It can be observed that the classification accuracy obtained by k-fold cross-validation procedure is comparably similar to the classification accuracy obtained in Tables 2-4.

Table A1: k-fold cross-validation using improved RLDA and four distinct classifiers on the SRBCT dataset.

Fold	J4.8	Naïve bayes	kNN	SVM pairwise
$k = 5$	80%	89%	92%	100%
$k = 10$	88%	92%	95%	100%

Table A2: k-fold cross-validation using improved RLDA and four distinct classifiers on the MLL dataset.

Fold	J4.8	Naïve bayes	kNN	SVM pairwise
$k = 5$	91%	94%	94%	95%
$k = 10$	87%	93%	95%	97%

Table A3: k-fold cross-validation using improved RLDA and four distinct classifiers on the Acute Leukemia dataset.

Fold	J4.8	Naïve bayes	kNN	SVM pairwise
$k = 5$	91%	97%	87%	94%
$k = 10$	87%	100%	95%	98%

Appendix II

In this appendix, we compare different values of regularization parameter with the proposed improved RLDA technique. In order to show this, we computed classification accuracy on four different values of α for RLDA technique. These are $\delta = [0.001, 0.01, 0.1, 1]$, where $\alpha = \delta * \lambda_W$ and λ_W is the maximum eigenvalue of within-class scatter matrix. We applied 3-fold cross-validation procedure on a number of datasets and shown the results in columns 2-5 of Table A2. The last column of the table denotes the classification accuracy using improved RLDA technique.

Table A4: Classification accuracy (in percentage) of RLDA and improved RLDA. The highest classification accuracies obtained are depicted in bold fonts.

Database	$\boldsymbol{\delta = 0.001}$	$\boldsymbol{\delta = 0.01}$	$\boldsymbol{\delta = 0.1}$	$\boldsymbol{\delta = 1}$	Improved RLDA
Acute Leukemia	98.6	98.6	98.6	**100**	**100.0**
MLL	95.7	95.7	95.7	95.7	**100.0**
SRBCT	**100.0**	**100.0**	**100.0**	96.2	**100.0**

It can be observed from the table that the different values of the regularization parameter give different classification accuracies and therefore, the choice of the regularization parameter affects the classification performance. Thus, it is important to select the regularization parameter correctly to get the good classification performance. It can be observed that for all the datasets, the proposed technique is exhibiting promising results.

Appendix III

Corollary 1: The value of regularization parameter is non-negative; i.e., $\alpha \geq 0$ for $r_w \leq r_t$, where $r_t = rank(\mathbf{S}_T)$ and $r_w = rank(\mathbf{S}_W)$.

Proof 1: From equation 2, we can write

$$J = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\mathbf{S}_W+\alpha\mathbf{I})\mathbf{w}}, \qquad \text{A1}$$

where $\mathbf{S}_B \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{S}_W \in \mathbb{R}^{r_t \times r_t}$. We can rearrange the above expression as

$$\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w} = J\mathbf{w}^{\mathrm{T}}(\mathbf{S}_W + \alpha\mathbf{I})\mathbf{w} \qquad \text{A2}$$

The eigenvalue decomposition (EVD) of $\mathbf{S}_W$ matrix (assuming $r_w < r_t$) can be given as $\mathbf{S}_W = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{r_t \times r_t}$ is an orthogonal matrix, $\mathbf{\Lambda}^2 = \begin{bmatrix} \mathbf{\Lambda}_w^2 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{\Lambda}_w = diag(q_1^2, q_2^2, \dots, q_{r_w}^2) \in \mathbb{R}^{r_w \times r_w}$ are diagonal matrices (as $r_w < r_t$). The eigenvalues $q_k^2 > 0$ for $k = 1,2,\dots,r_w$. Therefore,

$\mathbf{S}'_W = (\mathbf{S}_W + \alpha\mathbf{I}) = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{D} = \mathbf{\Lambda}^2 + \alpha\mathbf{I}$

or $\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}'_W\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{I}$ $\qquad$ A3

The between class scatter matrix $\mathbf{S}_B$ can be transformed by multiplying $\mathbf{U}\mathbf{D}^{-1/2}$ on the right side and $\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}$ on the left side of $\mathbf{S}_B$ as $\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_B\mathbf{U}\mathbf{D}^{-1/2}$. The EVD of this matrix will give

$$\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_B\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{E}\mathbf{D}_B\mathbf{E}^{\mathrm{T}}, \qquad \text{A4}$$

where $\mathbf{E} \in \mathbb{R}^{r_t \times r_t}$ is an orthogonal matrix and $\mathbf{D}_B \in \mathbb{R}^{r_t \times r_t}$ is a diagonal matrix. Equation A4 can be rearranged as

$$\mathbf{E}^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_B\mathbf{U}\mathbf{D}^{-1/2}\mathbf{E} = \mathbf{D}_B, \qquad \text{A5}$$

Let the leading eigenvalue of $\mathbf{D}_B$ is γ and its corresponding eigenvector is $\mathbf{e} \in \mathbf{E}$. Then equation A5 can be rewritten as

$$\mathbf{e}^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_B\mathbf{U}\mathbf{D}^{-1/2}\mathbf{e} = \gamma, \qquad \text{A6}$$

The eigenvector $\mathbf{e}$ can be multiplied right side and $\mathbf{e}^{\mathrm{T}}$ on left side of equation A3, we get

$$\mathbf{e}^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}'_W\mathbf{U}\mathbf{D}^{-1/2}\mathbf{e} = 1 \qquad \text{A7}$$

It can be seen from equations A3 and A5 that matrix $\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{E}$ diagonalizes both $\mathbf{S}_B$ and $\mathbf{S}'_W$, simultaneously. Also vector $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ simultaneously gives γ and unity eigenvalue in equations A6 and A7. Therefore, $\mathbf{w}$ is a solution of equation A2. Substituting $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ in equation A2, we get

$J = \gamma$; i.e., $\mathbf{w}$ is a solution of equation A2.

From Lemma 1, the maximum eigenvalue of expression $(\mathbf{S}_W + \alpha\mathbf{I})^{-\mathbf{1}}\mathbf{S}_B\mathbf{w} = \gamma\mathbf{w}$ is $\gamma_m = \lambda_{max} > 0$ (i.e., real, positive and finite). Therefore, the eigenvectors corresponding to this positive γ_m should also be in real hyperplane (i.e., the components of the vector $\mathbf{w}$ have to have real values). Since $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ with $\mathbf{w}$ to be in real hyperplane, we must have $\mathbf{D}^{-1/2}$ to be real.

Since $\mathbf{D} = \mathbf{\Lambda}^2 + \alpha\mathbf{I} = diag(q_1^2 + \alpha, q_2^2 + \alpha, \dots, q_{r_w}^2 + \alpha, \alpha, \dots, \alpha)$, we have

$\mathbf{D}^{-1/2} = diag(1/\sqrt{q_1^2 + \alpha}, 1/\sqrt{q_2^2 + \alpha}, \dots, 1/\sqrt{q_{r_w}^2 + \alpha}, 1/\sqrt{\alpha}, \dots, 1/\sqrt{\alpha})$.

Therefore, the elements of $\mathbf{D}^{-1/2}$, must satisfy $1/\sqrt{q_k^2 + \alpha} > 0$ and $1/\sqrt{\alpha} > 0$ for $k = 1,2, \dots, r_w$ (note $r_w < r_t$); i.e., α cannot be negative or $\alpha > 0$. Furthermore, if $r_w = r_t$ then matrix $\mathbf{S}_W$ will be a non-singular matrix and its inverse will exist. In this case, regularization is not required and therefore $\alpha = 0$. Thus, $\alpha \geq 0$ for $r_w \leq r_t$. This concludes the proof.

Reference

Anton, H., “Calculus”, *John Wiley and Sons*, New York, 1995.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and Korsemeyer, S.J., “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia”, *Nature Genetics*, vol. 30, pp 41-47, 2002.
[Data Source1: http://sdmc.lit.org.sg/GEDatasets/Datasets.html]
[Data Source2: http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63]

Banerjee M., Mitra S., Banka H., Evolutinary-rough feature selection in gene expression data, *IEEE Transaction on Systems, Man, and Cybernetics, Part C:*

Application and Reviews, vol. 37, 622–632, 2007.

Cong G., Tan K.-L., Tung A.K.H., Xu X., Mining top-k covering rule groups for gene expression data. In: *the ACM SIGMOD International Conference on Management of Data,* pp. 670-681, 2005.

Dai D.Q. and Yuen, P.C., “Regularized discriminant analysis and its application to face recognition”, *Pattern Recognition*, vol. 36, no. 3, pp. 845-847, 2003.

Dai D.Q., and Yuen, P.C., “Face recognition by regularized discriminant analysis”, *IEEE Transactions of SMC*, vol. 37, issue 4, pp. 1080-1085, 2007.

Ding, C. and Peng, H., "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". In Journal of Bioinformatics and Computer Biology, pp. 523-529, 2003.

Duda, R.O. and Hart, P.E., Pattern classification and scene analysis, Wiley, New York, 1973.

Dudoit,S., Fridlyand, J. and Speed, T.P, “Comparison of discriminant methods for the classification of tumors using gene expression data”, *Journal of the American Statistical Association*, vol. 97, pp. 77–87, 2002.

Friedman, J.H., “Regularized discriminant analysis”, *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, 1989.

Fukunaga, K., Introduction to statistical pattern recognition. *Academic Press Inc., Hartcourt Brace Jovanovich, Publishers.* 1990.

Furey T.S., Cristianini N., Duffy N., Bednarski D.W., Schummer M., Haussler D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* , vol. 16, no. 10, pp. 906-914, 2000.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander E.S., Molecular classification of cancer: class discovery and class prediction by gene

expression monitoring. *Science*, vol. 286, 531-537, 1999.
[Data Source: http://datam.i2r.a-star.edu.sg/datasets/krbd/]

Guo, Y., Hastie, T. and Tibshirani, R., 'Regularized discriminant analysis and its application in microarrays', *Biostatistics*, vol. 8, no. 1, pp. 86-100, 2007.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., "Gene Selection for Cancer Classification using Support Vector Machines". In: Machine Learning, vol. 46, pp. 389-422, 2002.

Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning, Springer, NY, USA, 2001.

Huang, R., Liu, Q., Lu, H. and Ma, S. "Solving the Small Sample Size Problem of LDA", *Proceedings of ICPR*, vol. 3, pp. 29-32, 2002.

Huang, Y., Xu, D., Nie, F., Semi-supervised dimension reduction using trace ratio criterion, IEEE Trans. Neural Networks and Learning Systems, vol. 23, no. 3, pp. 519-526, 2012a.

Huang, Y., Xu, D., Nie, F., Patch Distribution Compatible Semi-Supervised Dimension Reduction for Face and Human Gait Recognition," IEEE Trans. on Circuits and Systems for Video Technology, vol. 22, no. 3, pp. 479-488, 2012b.

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

Li J., Wong L., Using rules to analyse bio-medical data: a comparison between C4.5 and PCL, In: *Advances in Web-Age Information Management.* Berlin / Heidelberg: Springer, pp. 254-265, 2003.

Liu, J., Chen, S.C., Tan, X.Y., Efficient pseudo-inverse linear discriminant analysis and its nonlinear form for face recognition, Int. J. Patt. Recogn. Artif. Intell. vol. 21, no. 8, pp.

1265-1278, 2007.

Nie, F., Huang, H., Cai X., Ding, C., Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, NIPS, 2010.

Pan, W., "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments", *Bioinformatics*, vol. 18, pp. 546-554, 2002.

Pavlidis, P., Weston, J., Cai, J. and Grundy, W.N., "Gene functional classification from heterogeneous data", *International Conference on Computational Biology*, pp. 249-255, 2001.

Peng, H., Long, F. and Dong, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

Saeys, Y., Inza, I. and Larrañaga, P., "A review of feature selection techniques in bioinformatics". Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

Sharma, A., Imoto, S., and Miyano, S., "A top-r feature selection algorithm for microarray gene expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764, 2012a.

Sharma, A., Imoto, S., Miyano, S., A between-class overlapping filter-based method for transcriptome data analysis, Journal of Bioinformatics and Computational Biology, vol. 10, no. 5, pp. 1250010-1 1250010-20, 2012b.

Sharma, A. Imoto, S., Miyano, S., Sharma, V., "Null space based feature selection method for gene expression data", International Journal of Machine Learning and Cybernetics, vol. 3, issue 4, pp. 269-276, 2012c, DOI 10.1007/s13042-011-0061-9

Sharma, A., Koh, C.H., Imoto, S., and Miyano, S., Strategy of finding optimal number of features on gene expression data, Electronics Letters, IEE, vol. 47, no. 8, pp. 480-482, 2011.

Sharma, A., Paliwal, K.K., Fast principal component analysis using fixed-point algorithm, Pattern Recognition Letters, vol. 28, issue 10, pp. 1151-1155, 2007.

Sharma, A., Paliwal, K.K., “Rotational Linear Discriminant Analysis for Dimensionality Reduction”, IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 10, pp. 1336-1347, 2008a.

Sharma, A., Paliwal, K.K., A gradient linear discriminant analysis for small sample sized problem, Neural Processing Letters, vol. 27, no. 1, pp 17-24, 2008b.

Sharma, A., Paliwal, K.K., A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices, Pattern Recognition, vol. 45, pp. 2205-2213, 2012.

Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., “A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition”, Journal of Theoretical Biology, vol. 320, no. 7, pp. 41-46, 2013a.

Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., Sharma, V., Ananthanarayanan, R. A Feature Selection Method using Fixed-Point Algorithm for DNA microarray gene expression data, International Journal of Knowledge Based and Intelligent Engineering Systems, 2013b (accepted).

Su, Y., Murali, T.M., Pavlovic, V. and Kasif, S., RankGene: identification of diagnostic genes based on expression data, *Bioinformatics*, pp. 1578–1579, 2003.

Tan A.C., Gilbert D., Ensemble machine learning on gene expression data for cancer classification, *Appl. Bioinformatics*, 2(3 Suppl), pp. S75-83, 2003.

Tao L., Zhang, C. and Ogihara, M., “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression”, *Bioinformatics*, vol, 20, no. 14, pp. 2429-2437, 2004.

Thomas, J., Olson, J.M., Tapscott, S.J. and Zhao, L.P., “An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles", *Genome Research*, vol. 11, pp. 1227–1236, 2001.

Tibshirani R, Regression shrinkage and selection via the lasso, J Royal Stat Soc B, vol. 58, no. 1, pp. 267-288, 1996.

Wang, A. and Gehan, E.A., "Gene selection for microarray data analysis using principal component analysis", *Statistics in Medicine*, vol. 24, pp. 2069-2087, 2005.

Wu, G., Xu, W., Zhang, Y., Wei, Y., "A preconditioned conjugate gradient algorithm fo GeneRank with application to microarray data mining", Data Mining and Knowledge Discovery, 2011, DOI: 10.1007/s10618-011-0245-7.

Xu, D., Yan, S., Semi-supervised bilinear subspace learning, IEEE Trans. Image Processing, vol., 18, no. 7, pp. 1671-1676, 2009.

Zhou, L., Wang, L., Shen, C., Barnes, N., Hippocampal shape classification using redundancy constrained feature selection, Medical Image Computing and Computer-Assisted Intervention, MICCAI 2010, Lecture Notes in Computer Science, vol. 6362, pp 266-273, 2010.