Journal of Experimental & Clinical Cancer Research

BioMed Central

Research

Open Access

Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data

Desheng Huang[1,2], Yu Quan[3], Miao He[4] and Baosen Zhou*[2,5]

Addresses: [1]Department of Mathematics, College of Basic Medical Sciences, China Medical University, Shenyang 110001, China, [2]Key Laboratory of Cancer Etiology and Intervention, University of Liaoning Province, China, [3]Computer Center, Affiliated Shengjing Hospital, China Medical University, Shenyang 110004, China, [4]Information Center, the First Affiliated Hospital, China Medical University, Shenyang 110001, China and [5]Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110001, China

E-mail: Desheng Huang - dshuang@mail.cmu.edu.cn; Yu Quan - quanyu@sj-hospital.org; Miao He - job-mail@263.net; Baosen Zhou* - bszhou@mail.cmu.edu.cn
*Corresponding author

Published: 10 December 2009

Journal of Experimental & Clinical Cancer Research 2009, **28**:149 doi: 10.1186/1756-9966-28-149

Received: 8 November 2009
Accepted: 10 December 2009

This article is available from: http://www.jeccr.com/content/28/1/149

Abstract

Background: More studies based on gene expression data have been reported in great detail, however, one major challenge for the methodologists is the choice of classification methods. The main purpose of this research was to compare the performance of linear discriminant analysis (LDA) and its modification methods for the classification of cancer based on gene expression data.

Methods: The classification performance of linear discriminant analysis (LDA) and its modification methods was evaluated by applying these methods to six public cancer gene expression datasets. These methods included linear discriminant analysis (LDA), prediction analysis for microarrays (PAM), shrinkage centroid regularized discriminant analysis (SCRDA), shrinkage linear discriminant analysis (SLDA) and shrinkage diagonal discriminant analysis (SDDA). The procedures were performed by software R 2.80.

Results: PAM picked out fewer feature genes than other methods from most datasets except from Brain dataset. For the two methods of shrinkage discriminant analysis, SLDA selected more genes than SDDA from most datasets except from 2-class lung cancer dataset. When comparing SLDA with SCRDA, SLDA selected more genes than SCRDA from 2-class lung cancer, SRBCT and Brain dataset, the result was opposite for the rest datasets. The average test error of LDA modification methods was lower than LDA method.

Conclusions: The classification performance of LDA modification methods was superior to that of traditional LDA with respect to the average error and there was no significant difference between theses modification methods.

Background

Conventional diagnosis of cancer has been based on the examination of the morphological appearance of stained tissue specimens in the light microscope, which is subjective and depends on highly trained pathologists. Thus, the diagnostic problems may occur due to inter-observer variability. Microarrays offer the hope that cancer classification can be objective and accurate. DNA microarrays measure thousands to millions of gene expressions at the same time, which could provide the clinicians with the information to choose the most appropriate forms of treatment.

Studies on the diagnosis of cancer based on gene expression data have been reported in great detail, however, one major challenge for the methodologists is the choice of classification methods. Proposals to solve this problem have utilized many innovations including the introduction of sophisticated algorithms for support vector machines [1] and the proposal of ensemble methods such as random forests [2]. The conceptually simple approach of linear discriminant analysis (LDA) and its sibling, diagonal discriminant analysis (DDA) [3-5], remain among the most effective procedures also in the domain of high-dimensional prediction. In the present study, our main focus will be solely put on the LDA part and henceforth the term "discriminant analysis" will stand for the meaning of LDA unless otherwise emphasized. The traditional way of doing discriminant analysis is introduced by R. Fisher, known as the linear discriminant analysis (LDA). Recently some modification of LDA have been advanced and gotten good performance, such as prediction analysis for microarrays (PAM), shrinkage centroid regularized discriminant analysis(SCRDA), shrinkage linear discriminant analysis(SLDA) and shrinkage diagonal discriminant analysis(SDDA). So, the main purpose of this research was to describe the performance of LDA and its modification methods for the classification of cancer based on gene expression data.

Cancer is not a single disease, there are many different kinds of cancer, arising in different organs and tissues through the accumulated mutation of multiple genes. Many previous studies only focused on one method or single dataset and gene selection is much more difficult in multi-class situations [6,7]. Evaluation of the most commonly employed methods may give more accurate results if it is based on the collection of multiple databases from the statistical point of view.

In summary, we investigate the performance of LDA and its modification methods for the classification of cancer based on multiple gene expression datasets.

Methods

Procedure for the classification of cancer is shown as follows. First, a classifier is trained on a subset (training set) of gene expression dataset. Then, the mature classifier is used for unknown subset (test set) and predicting each observation's class. The detailed information about classification procedure is shown in Figure 1.

Datasets

Six publicly available microarray datasets [8-14] were used to test the above described methods and we call them 2-class lung cancer, colon, prostate, multi-class lung cancer, SRBCT and brain following the naming there. Due to the fact that microarray-based studies may report findings that are not reproducible, after reviewing literature we selected these above public datasets with the consideration of our research topic and cross-comparison with other similar studies. The main features of these datasets are summarized in Table 1.

Data pre-processing

To avoid the noise of the dataset, pre-processing was necessary in the analysis. Absolute transformation was first performed on the original data. The data was transformed to have a mean of 0 and standard deviation

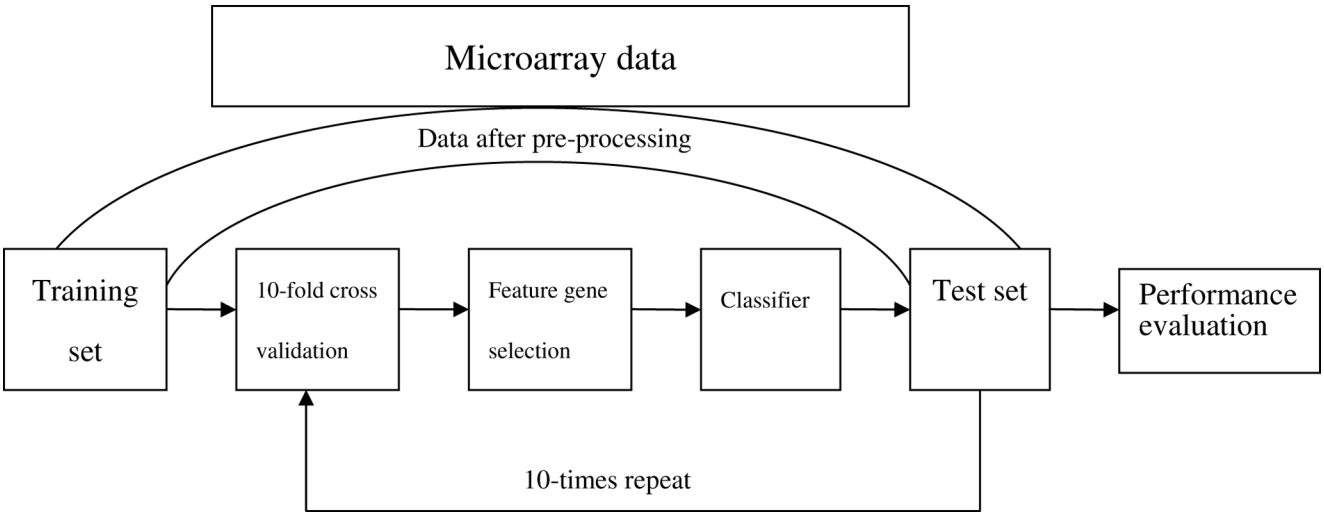

Figure 1
Framework for the procedure of classification.

Table 1: Characteristics of the six microarray datasets used

Dataset	No. of samples	Classes (No. of samples)	No. of genes	Original ref.	Website
Two-class lung cancer	181	MPM(31), adenocarcinoma(150)	12533	[8]	http://www.chestsurg.org
Colon	62	normal(22), tumor(40)	2000	[9]	http://microarray.princeton.edu/oncology/affydata/index.html
Prostate	102	normal(50), tumor(52)	6033	[10]	http://microarray.princeton.edu/oncology/affydata/index.html
Multi-class lung cancer	68(66) [a]	adenocarcinoma(37), combined(1), normal(5), small cell(4), squamous cell(10), fetal(1), large cell(4), lymph node(6)	3171	[11,12]	http://www-genome.wi.mit.edu/mpr/lung/
SRBCT	88(83) [b]	Burkitt lymphoma (29), Ewing sarcoma (11), neuroblastoma (18), rhabdomyosarcoma (25), non-SRBCTs(5)	2308	[13]	http://research.nhgri.nih.gov/microarray/Supplement/
Brain	42(38) [c]	medulloblastomas(10), CNS AT/RTs(5), rhabdoid renal and extrarenal rhabdoid tumours(5), supratentorial PNETs(8), non-embryonal brain tumours (malignant glioma) (10), normal human cerebella(4)	5597	[14]	http://research.nhgri.nih.gov/microarray/Supplement/

Note: Some samples were removed for keeping adequate number of each type.
a. One combined and one fetal cancer samples were removed, and real sample size is 66;
b. Five non-SRBCT samples were removed, and real sample size is 83;
c. Four normal tissue samples were removed, and real sample size is 38.

of 1 after logarithmic transformation and normalization. When the original data had already experienced the above transformation, it entered next step directly.

Algorithms for feature gene selection

Notation

Let x_{ij} be the expression level of gene j in the sample i, and y_i be the cancer type for sample i, j = 1,...,p and response $y_i \in \{1,...,K\}$. Denote $Y = (y_1,...,y_n)^T$ and $x_i = (x_{i1},...,x_{ip})^T$, i = 1,...,n. Gene expression data on p genes for n mRNA samples may be summarized by an n × p matrix $X = (x_{ij})_{n \times p}$. Let C_k be indices of the n_k samples in class k, where n_k denotes the number of observations belonging to class k, $n = n_1+...+n_K$. A predictor or classifier for K tumor classes can be built from a learning set L by C(.,L); the predicted class for an observation x* is C(x*,L). The jth component of the 1centroid for class k is $\bar{x}_{kj} = \sum_{i \in C_k n} x_{ij} / n_k$, the jth component of the overall centroid is $\bar{x}_j = \sum_{i=1} x_{ij} / n$.

Prediction analysis for microarrays/nearest shrunken centroid method, PAM/NSC

PAM [3] algorithm tries to shrink the class centroids ($\bar{x}_{kj}$) towards the overall centroid $\bar{x}_j$.

$$\text{Let } d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k \cdot (s_j + s_0)} \quad (1)$$

where d_{kj} is a t statistic for gene j, comparing class k to the overall centroid, and s_j is the pooled within-class standard deviation for gene j:

$$s_j^2 = \frac{1}{n-K} \sum_k \sum_{i \in C_k} (x_{ij} - \bar{x}_{kj}) \quad (2)$$

and $m_k = \sqrt{1/n_k + 1/n}$, s_0 is a positive constant and usually equal to the median value of the s_j over the set of genes.

Equation(1) can be transformed to

$$\bar{x}_{kj} = \bar{x}_j + m_k (s_j + s_0) d_{kj} \quad (3)$$

PAM method shrinks each d_{kj} toward zero, and giving d'_{kj} yielding shrunken centroids

$$\bar{x}'_{kj} = \bar{x}_j + m_k (s_j + s_0) d'_{kj} \quad (4)$$

Soft thresholding is defined by

$$d'_{kj} = sign(d_{kj})(|d_{kj}| - \Delta)_+ \quad (5)$$

where + means positive part ($t_+ = t$ if $t > 0$ and zero otherwise). For a gene j, if d_{kj} is shrunken to zero for all classes k, then the centroid for gene j is $\bar{x}_j$, the same for all classes. Thus gene j does not contribute to the

nearest-centroid computation. Soft threshold Δ was chosen by cross-validation.

Shrinkage discriminant analysis, SDA

In SDA, Feature selection is controlled using higher criticism threshold (HCT) or false non-discovery rates (FNDR) [5]. The HCT is the order statistic of the Z-score corresponding to index i maximizing $(\frac{i}{p} - \pi(i))/\sqrt{\frac{i}{p}(1-\frac{i}{p})}$, π_i is the p-value associated with the ith Z-score and $\pi_{(i)}$ is the ith order statistic of the collection of p-values($1 \le i \le p$). The ideal threshold optimizes the classification error. SDA consists of Shrinkage linear discriminant analysis (SLDA) and Shrinkage diagonal discriminant analysis (SDDA) [15,16].

Shrunken centroids regularized discriminant analysis, SCRDA

There are two parameters in SCRDA [4], one is α ($0<\alpha<1$), the other is soft threshold Δ. The choosing the optimal tuning parameter pairs (α, Δ) is based on cross-validation. A "Min-Min" rule was followed to identify the optimal parameter pair (α, Δ):

First, all the pairs (α, Δ) that corresponded to the minimal cross-validation error from training samples were found.

Second, the pair or pairs that used the minimal number of genes were selected.

When there was more than one optimal pair, the average test error based on all the pairs chosen would be calculated. As traditional LDA is not suitable to deal with the "large p, small N" paradigm, so we did not adopt it to select feature genes.

Algorithms of LDA and its modification methods for classification

Linear discriminant analysis, LDA

Fisher linear discriminant analysis (FLDA, or for short, LDA) [17] projects high dimension data x into one dimension axle to find linear combinations **xa** with large ratios of between-group to within-group sums of squares. Fisher's criteria can be defined as:

$$\max \frac{a'Ba}{a'Wa} \tag{6}$$

Where B and W denote the matrices of between-group and within-group sums of squares and cross-products.

Class k sample means $\bar{x}_k = (\bar{x}_{k1}, ..., \bar{x}_{kp})$ can be gotten from learning set L, and for a new tumor sample with gene expression x*, the predicted class for x* is the class whose mean vector $\bar{x}_k$ is closest to x* in the space of discriminant variables, that is

$$C(x^*, L) = \arg\min_k d_k(x^*) \tag{7}$$

where $d_k^2(x^*) = \sum_{l=1}^{s}[(x^* - \bar{x}_k)v_l]^2$, v_l is eigenvector, s is the number of feature genes.

When numbers of classes K = 2, FLDA yields the same classifier as the maximum likelihood (ML) discriminant rule for multivariate normal class densities with the same covariance matrix.

Prediction analysis for microarrays/nearest shrunken centroid method, PAM/NSC

PAM [3] assumes that genes are independent, the target classes correspond to individual (single) clusters and classify test samples to the nearest shrunken centroid, again standardizing by $s_j + s_0$. The relative number of samples in each class is corrected at the same time. For a test sample (a vector) with expression levels x^*, the discriminant score for class k was defined by,

$$\delta_k(x^*) = \sum_{j=1}^{P} \frac{(x^* - x_{kj}')^2}{(s_j + s_0)^2} - 2\log(\pi_k) \tag{8}$$

where $\pi_k = n_k/n$ or $\pi_k = 1/K$ is class prior probability, $\sum_{k=1}^{K}\pi_k = 1$. This prior probability gives the overall frequency of class k in the population. The classification rule is

$$x^* \in k,\ (k = \underset{k'}{\operatorname{argmin}}\,\delta_{k'}(x^*) = \underset{k'}{\operatorname{argmin}}(x^* - x_{k'}')^T D^{-1}(x^* - \bar{x}_{k'}') - \log\pi_{k'}) \tag{9}$$

Here $\hat{D}$ was the diagonal matrix taking the diagonal elements of $\hat{\Sigma}$. If the smallest distances are close and hence ambiguous, the prior correction gives a preference for larger classes, because they potentially account for more errors.

Shrinkage discriminant analysis, SDA

The corresponding discriminant score [5] was defined by

$$\Delta_k^{LDA}(x^*) = \omega_k^T \delta_k(x^*) + \log(\pi_k) \tag{10}$$

Where $\omega_k = P^{-\frac{1}{2}}V^{-\frac{1}{2}}(\bar{x}_k - \bar{x})$, $\delta_k = P^{-\frac{1}{2}}V^{-\frac{1}{2}}(x^* - \frac{\bar{x}_k + \bar{x}}{2})$, $V = diag\{\sigma_1^2, \cdots, \sigma_p^2\}$, $P = (\rho_{ij})$ and $\bar{x} = \sum_{j=1}^{k}\frac{n_j}{n}\bar{x}_j$

Algorithm of SCRDA

A new test sample was classified by regularized discriminant function [4],

$$\tilde{d}_k(x^*) = (x^*)^T \tilde{\Sigma}^{-1} \overline{x}_k - \frac{1}{2} \overline{x}_k^T \tilde{\Sigma}^{-1} \overline{x}_k + \log \pi_k \qquad (11)$$

Covariance was estimated by

$$\tilde{\Sigma} = \alpha \hat{\Sigma} + (1-\alpha) I_P \qquad (12)$$

where $0 \le \alpha \le 1$

In the same way, sample correlation matrix was substituted by $\hat{R} = \hat{D}^{-\frac{1}{2}} \hat{\Sigma} D^{-\frac{1}{2}}$ $\tilde{R} = \alpha \hat{R} + (1-\alpha) I_P$.

Then the regularized sample covariance matrix was computed by $\tilde{\Sigma} = \hat{D}^{-\frac{1}{2}} \tilde{R} D^{\frac{1}{2}}$.

Study design and program realization

We used 10-fold cross-validation (CV) to divide the pre-processed dataset into 10 approximately equal-size parts by random sampling. It worked as follows: we fit the model on 90% of the samples and then predicted the class labels of the remaining 10% (the test samples). This procedure was repeated 10 times to avoid overlapping test sets, with each part playing the role of the test samples and the errors on all 10 parts added together to compute the overall error [18]. R software (version 2.80) with packages MASS, pamr, RDA, SDA was used for the realization of the above described methods [19]. A tolerance value was set to decide if a matrix is singular. If variable had within-group variance less than tol^2, LDA fitting iteration would stop and report the variable as constant. In practice, we set a very small tolerance value 1×10^{-14}, and no singular was detected.

Results

Feature genes selection

As shown in Table 2, PAM picked out fewer feature genes than other methods from most datasets except from Brain dataset. For the two methods of shrinkage discriminant analysis, SLDA selected more genes than SDDA from most datasets except from 2-class lung cancer dataset. When comparing SLDA with SCRDA, SLDA selected more genes than SCRDA from 2-class lung cancer, SRBCT and Brain dataset, the result was opposite for the rest datasets.

Table 2: Numbers of feature genes selected by 4 methods for each dataset

Dataset	PAM	SDDA	SLDA	SCRDA
2-class lung cancer	7.98	422.74	407.83	118.72
Colon	25.72	65.67	117.08	214.87
Prostate	83.13	120.53	187.91	217.47
Multi-class lung cancer	45.26	57.98	97.27	1015.00
SRBCT	30.87	114.32	131.24	86.22
Brain	69.11	115.04	182.01	26.83

Performance comparison for methods based on different datasets

The performance of the methods described above was compared by average test error using 10-fold cross validation. We ran 10 cycles of 10-fold cross validation. The average test errors were calculated based on the incorrectness of the classification of each testing samples. For example, for the 2-class lung cancer dataset, using the LDA method based on PAM as the feature gene method, 30 samples out of 100 sample test sets were incorrectly classified, resulting in an average test error of 0.30.

The significance of the performance difference between these methods was judged depending on whether or not their 95% confidence intervals of accuracy overlapped. Here, if the upper limit was greater than 100%, it was treated as 100%. If two methods had non-overlapping confidence intervals, their performances were significantly different. The bold fonts in Table 3 shows the performances of PAM, SDDA, SLDA and SCRDA, when they were used both for feature gene selection and classification. As shown in Table 3, the performance of LDA modification methods is superior to traditional LDA method, while there is no significant difference between theses modification methods (Figure 2).

Discussion

Microarrays are capable of determining the expression levels of thousands of genes simultaneously and hold great promise to facilitate the discovery of new biological knowledge [20]. One feature of microarray data is that the number of variables *p* (genes) far exceeds the number of samples *N*. In statistical terms, it is called 'large *p*, small *N*' problem. Standard statistical methods in classification do not work well or even at all, so improvement or modification of existing statistical methods is needed to prevent over-fitting and produce more reliable estimations. Some ad-hoc shrinkage methods have been proposed to utilize the shrinkage ideas and prove to be useful in empirical studies [21-23]. Distinguishing normal samples from tumor samples is essential for successful diagnosis or treatment of cancer. And, another important problem is in characterizing multiple types of tumors. The problem of multiple classifications has recently received more attention in the context of DNA microarrays. In the present study, we first presented an evaluation of the performance of LDA and its modification methods for classification with 6 public microarray datasets.

The gene selection method [6,24,25], the number of selected genes and the classification method are three critical issues for the performance of a sample classification. Feature selection techniques can be organized into

Table 3: Average test error of LDA and its modification methods (10 cycles of 10-fold cross validation)

Dataset	Gene selection methods	Performance				
		LDA	PAM	SDDA	SLDA	SCRDA
2-class Lung cancer data(n = 181, p = 12533, K = 2)	PAM	0.30	**0.26**	0.15	0.16	0.42
	SDDA	0.17	0.11	**0.1**	0.11	0.1
	SLDA	0.47	0.3	0.3	**0.3**	0.32
	SCRDA	0.73	0.20	0.19	0.17	**0.19**
Colon data(n = 62, p = 2000, K = 2)	PAM	1.30	**0.82**	0.8	0.86	0.86
	SDDA	2.25	2.09	**1.33**	1.29	1.25
	SLDA	1.12	0.74	0.75	**0.77**	0.80
	SCRDA	1.19	0.77	0.77	0.75	**0.78**
Prostate data(n = 102, p = 6033, K = 2)	PAM	2.87	**0.89**	0.82	0.81	1.00
	SDDA	2.53	0.71	**0.72**	0.68	0.74
	SLDA	1.75	0.7	0.64	**0.64**	0.70
	SCRDA	2.15	0.57	0.59	0.57	**0.61**
Multi-class lung cancer data(n = 66, p = 3171, K = 6)	PAM	2.13	**1.16**	1.21	1.28	1.19
	SDDA	1.62	1.32	**1.32**	1.31	1.30
	SLDA	1.62	1.31	1.32	**1.26**	1.34
	SCRDA	1.63	1.43	1.45	1.58	**1.35**
SRBCT data(n = 83, p = 2308, K = 4)	PAM	0.17	**0.01**	0.01	0.03	0.01
	SDDA	2.45	0.03	**0.02**	0	0.03
	SLDA	2.87	0	0	**0**	0
	SCRDA	2.32	0.03	0.03	0.02	**0.03**
Brain data(n = 38, p = 5597, K = 4)	PAM	1.14	**0.57**	0.57	0.58	0.61
	SDDA	1.09	0.61	**0.62**	0.63	0.55
	SLDA	0.89	0.60	0.60	**0.57**	0.58
	SCRDA	0.84	0.56	0.54	0.54	**0.57**

three categories, filter methods, wrapper methods and embedded methods. LDA and its modification methods belong to wrapper methods which embed the model hypothesis search within the feature subset search. In the present study, different numbers of gene have been selected by different LDA modification methods. There is no theoretical estimation of the optimal number of selected genes and the optimal gene set can vary from data to data [26]. So we did not focus on the combination of the optimal gene set by one feature gene selection method and one classification algorithm. In this paper we just describe the performance of LDA and its modification methods under the same selection method in different microarray dataset.

Various statistical and machine learning methods have been used to analyze the high dimensional data for cancer classification. These methods have been shown to have statistical and clinical relevance in cancer detection for a variety of tumor types. In this study, it has been shown that LDA modification methods have better performance than traditional LDA under the same gene selection criterion. Dudoit also reported that simple classifiers such as DLDA and Nearest Neighbor performed remarkably well compared with more sophisticated ones, such as aggregated classification trees [27]. It indicates that LDA modification methods did a good job in some situations. Zhang *et al* [28] developed a fast algorithm of generalized linear discriminant analysis (GLDA) and applied it to seven public cancer datasets. Their study included 4 same datasets (Colon, Prostate, SRBCT and Brain) as those in our study and adopted a 3-fold cross-validation design. The average test errors of our study were less than those of their study, while there was no statistical significance of the difference. The results reported by Guo *et al* [4] are of concordance with ours except for the colon dataset. Their study also included the above mentioned 4 same datasets and they found that in the colon dataset the average test error of SCRDA was as same as PAM, while in the present study we found that the average test error of SCRDA was slightly less than that of PAM.

There are several interesting problems that remain to be addressed. A question is raised that when comparing the predictive performance of different classification methods on different microarray data, is there any difference between various methods, such as leave-one-out cross-validation and bootstrap [29,30]? And another interesting further step might be a pre-analysis of the data to

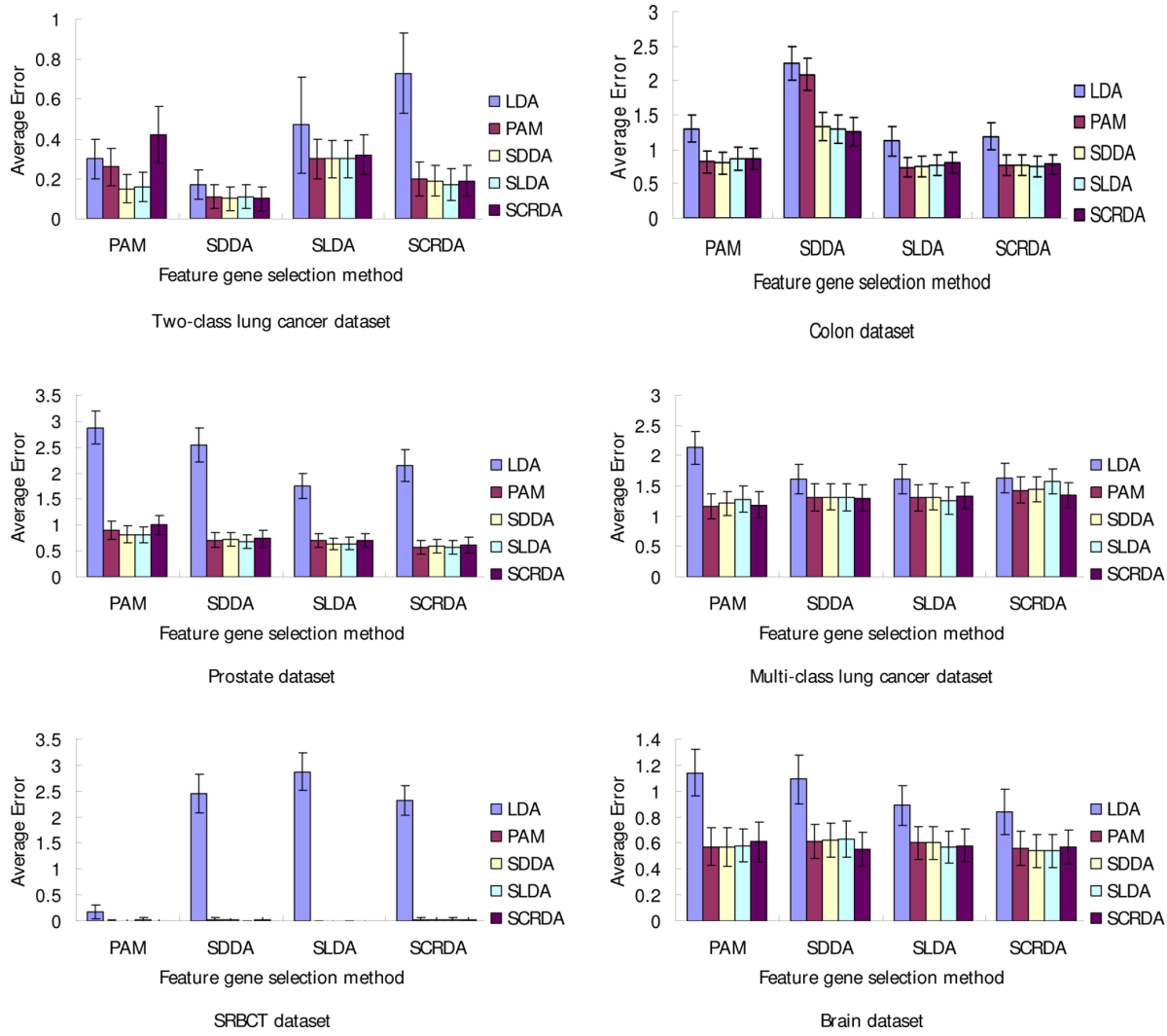

Figure 2
Comparison of classification performance for different datasets. The y-axis shows the average error and the x-axis indicates the gene selection methods: PAM, SDDA, SLDA and SCRDA. Error bars (± 1.96 SE) are provided for the classification methods.

choose a suitable gene selection method. Despite the great promise of discriminant analysis in the field of microarray technology, the complexity and the multiple choices of the available methods are quite difficult to the bench clinicians. This may influence the clinicians' adoption of microarray data based results when making decision on diagnosis or treatment. Microarray data's widespread clinical relevance and applicability still need to be resolved.

Conclusions

An extensive survey in building classification models from microarray data with LDA and its modification methods has been conducted in the present study. The study showed that the modification methods are superior to LDA in the prediction accuracy.

List of abbreviations

CV: Cross-validation; DDA: diagonal discriminant analysis; FNDR: False non-discovery rates; GLDA: generalized linear discriminant analysis; HCT: Higher criticism threshold; LDA: linear discriminant analysis; NSC: nearest shrunken centroid method; PAM: prediction analysis for microarrays; SCRDA: Shrinkage centroid regularized discriminant analysis; SDA: Shrinkage discriminant analysis; SDDA: Shrinkage diagonal discriminant analysis; SLDA: Shrinkage linear discriminant analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DH conceived the study and drafted the manuscript. DH and YQ performed the analyses. MH provided guidance and discussion on the methodology. BZ attracted partial funding and participated in the design of the analysis strategy. All authors read and approved the final version of this manuscript.

Acknowledgements

This study was partially supported by Provincial Education Department of Liaoning (No.2008S232), Natural Science Foundation of Liaoning province (No.20072103) and China Medical Board (No.00726.). The authors are most grateful to the contributors of the datasets and R statistical software. The authors thank the two reviewers for their insightful comments which led to an improved version of the manuscript.

References

1. Guyon I, Weston J, Barnhill and Vapnik V: **Gene Selection for Cancer Classification using Support Vector Machines.** *Mach Learn* 2002, **46:**389–422.
2. Breiman L: **Random Forests.** *Mach Learn* 2001, **45:**5–32.
3. Tusher VG, Tibshirani R and Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116–5121.
4. Guo Y, Hastie T and Tibshirani R: **Regularized linear discriminant analysis and its application in microarrays.** *Biostatistics* 2005, **8:**86–100.
5. Schäfer J and Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Stat Appl Genet Mol Biol* 2005, **4**.
6. Yeung KY, Bumgarner RE and Raftery AE: **Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data.** *Bioinformatics* 2005, **21:**2394–2402.
7. Li T, Zhang C and Ogihara M: **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.** *Bioinformatics* 2004, **20:**2429–2437.
8. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ and Bueno R: **Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma.** *Cancer Res* 2002, **62:**4963–4967.
9. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96:**6745–6750.
10. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR and Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1:**203–209.
11. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ and Meyerson M: **Classification of human lung carcinomas by mRNA expressionprofiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98:**13790–13795.
12. Parmigiani G, Garrett-Mayer ES, Anbazhagan R and Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer.** *Clin Cancer Res* 2004, **10:**2922–2927.
13. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C and Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7:**673–679.
14. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES and Golub TR: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415:**436–442.
15. Opgen-Rhein R and Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach.** *Stat Appl Genet Mol Biol* 2007, **6:**Article9.
16. Schäfer J and Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Stat Appl Genet Mol Biol* 2005, **4:**Article32.
17. Fisher RA: **The Use of Multiple Measurements in Taxonomic Problems.** *Annuals of Eugenics* 1936, **7:**179–188.
18. Hastie T, Tibshirani R and Friedman J: **The elements of statistical learning; data mining, inference and prediction.** New York: Springer; 2001, 193–224.
19. R Development Core Team R: **A language and environment forstatistical computing.** R Foundation for StatisticalComputing, Vienna, Austria; 2009 http://www.R-project.org, ISBN 3-900051-07-0.
20. Campioni M, Ambrogi V, Pompeo E, Citro G, Castelli M, Spugnini EP, Gatti A, Cardelli P, Lorenzon L, Baldi A and Mineo TC: **Identification of genes down-regulated during lung cancer progression: a cDNA array study.** *J Exp Clin Cancer Res* 2008, **27:**38.
21. Tusher VG, Tibshirani R and Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116–5121.
22. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J Royal Statist Soc B* 1996, **58:**267–288.
23. Xie Y, Pan W, Jeong KS and Khodursky A: **Incorporating prior information via shrinkage: a combined analysis of genome-wide location data and gene expression data.** *Stat Med* 2007, **26:**2258–2275.
24. Li Y, Campbell C and Tipping M: **Bayesian automatic relevance determination algorithms for classifying gene expression data.** *Bioinformatics* 2002, **18:**1332–1339.
25. Diaz-Uriarte R: **Supervised methods with genomic data: a review and cautionary view.** *Data analysis and visualization in genomics and proteomics.* Hoboken: John Wiley & Sons, Ltd: Francisco Azuaje, Joaquín Dopazo 2005, 193–214.
26. Tsai CA, Chen CH, Lee TC, Ho IC, Yang UC and Chen JJ: **Gene selection for sample classifications in microarray experiments.** *DNA Cell Biol* 2004, **23:**607–614.
27. Dudoit S, Fridlyand J and Speed TP: **Comparison of Discrimination Methods for the Classification o Tumors Using Gene Expression Data.** *J Am Stat Assoc* 2002, **97:**77–87.
28. Li H, Zhang K and Jiang T: **Robust and accurate cancer classification with gene expression profiling.** *Proc IEEE Comput Syst Bioinform Conf: 8-11 August 2005; California* 2005, 310–321.
29. Breiman L and Spector P: **Submodel selection and evaluation in regression: the x-random case.** *Int Stat Rev* 1992, **60:**291–319.
30. Efron B: **Bootstrap methods: Another look at the jackknife.** *Ann Stat* 1979, **7:**1–26.