Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 19 (2009) 1635–1641
Progress in Natural Science
www.elsevier.com/locate/pnsc

Sparse discriminant analysis for breast cancer biomarker identification and classification

Yu Shi [a,b], Daoqing Dai [a,*], Chaochun Liu [a,b], Hong Yan [b,c]

[a] Center for Computer Vision and Department of Mathematics, Sun Yat-Sen (Zhongshan) University, Guangzhou 510275, China
[b] Department of Electric Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China
[c] School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia

Received 23 February 2009; received in revised form 20 April 2009; accepted 21 April 2009

Abstract

Biomarker identification and cancer classification are two important procedures in microarray data analysis. We propose a novel unified method to carry out both tasks. We first preselect biomarker candidates by eliminating unrelated genes through the BSS/WSS ratio filter to reduce computational cost, and then use a sparse discriminant analysis method for simultaneous biomarker identification and cancer classification. Moreover, we give a mathematical justification about automatic biomarker identification. Experimental results show that the proposed method can identify key genes that have been verified in biochemical or biomedical research and classify the breast cancer type correctly.

Keywords: Biomarker identification; Cancer classification; Discriminant analysis; Maximum penalized likelihood; Microarray data analysis

1. Introduction

Microarray technology has emerged as a powerful tool allowing investigators to monitor the expression levels of thousands of genes simultaneously. It is now possible to identify tissue samples by their gene expression profiles such as the healthy or cancerous. Compared to the standard histopathological test, this technology for the diagnosis of cancer is more accurate, reliable and objective [1,2]. Many cases of hereditary breast cancer are due to mutations in either the BRCA1 or the BRCA2 gene. Hedenfalk et al. [3] proposed that the histopathological changes in these cancers are often characteristic of the mutant genes and they gave a framework to classify breast cancers associated with BRCA1 or BRCA2 mutation by the identified gene expression profiles. This was verified as a more powerful tool to predict the outcome of disease in patients with breast cancer than standard systems based on clinical and histological criteria [4].

In recent years, a number of statistical and machine-leaning methods have been used to identify tumor tissues with distinct patterns of gene expression. Gene selection and cancer classification are two highly correlated problems in the field of gene expression analysis of cancer samples, which are also called biomarker identification and cancer diagnosis. Gene selection requires the identification of several important genes which capture the information related to the cancer from thousands of genes. An effective gene selection should benefit classifier to design with better accuracy and interpretability. Cancer classification requires building a classifier from training data and predicting the cancer type of a new sample based on its gene expression.

Li [5] combined the genetic algorithm and the k-nearest neighbor (KNN) method to identify genes that can jointly discriminate different classes of samples. Tibshirani et al. [6] proposed an enhanced nearest centroid classifier by a shrunken centroid method to identify the subset of genes

* Corresponding author. Tel.: +86 20 8411 0141; fax: +86 20 8403 7978.
E-mail address: stsddq@mail.sysu.edu.cn (D. Dai).

1002-0071/$ - see front matter
doi:10.1016/j.pnsc.2009.04.013

that best characterize each class. Zhou et al. [7] proposed a mutual information-based gene and feature selection method, collaborating with the nonlinear probit regression approach for cancer classification. Yang et al. [8] proposed a feature ensemble method to identify critical disease-relevant genes based on their classifying performance. Dudoit et al. [9] compared the performance of some existing discrimination methods for cancer classification with the expression of genes selected based on the ratio of their between-groups to within-groups sum of squares (BSS/WSS). These methods include the nearest neighbor classifier, Fisher linear discriminant analysis, diagonal linear and quadratic discriminant analysis (DiagLDA, DiagQDA), classification trees, bagging and boosting. Focusing on breast cancers, Desper et al. [10] used a phylogenetic method to separate breast cancers to BRCA1-mutated, BRCA2-mutated, and sporadic. Lin [11] built a mixture model to explain progression pathways of heterogeneous breast tumors.

Many existing techniques consist of two steps: gene selection or feature extraction and cancer classification. Some of them undertake the two steps separately. Although they may be useful in practice, the methods that select marker genes independent of classifiers fail to take mutual information among different cancer types into account, and the methods that select genes based on individual genes fail to take mutual information among genes into account. Furthermore, the separation of these two tasks may occlude informative structure in the data.

Moreover, feature extraction methods establishing relations among genes provide a better performance for cancer classification. These methods include partial least squares [12], principal component analysis [13], generalized discriminant analysis [14] and uncorrelated discriminant analysis [15]. It is noticed that each feature extracted in these methods is a linear combination of all gene expressions. Consequently, the procedure ignores disease-related genes whose expression patterns are excellent diagnostic indicators, although they are biologically meaningful. In a review, Hilario [16] outlined approaches for dimensionality reduction in proteomic biomarker studies and suggested that combining different methods is a potential alternative to single-method selection. Li [17] proposed an ensemble decision approach to mining complex disease genes, and with these identified genes they achieved the best classification performance.

In this paper, we propose a novel unified approach for simultaneous biomarker identification and cancer classification. Linear discriminant analysis (LDA) is a widely used method for pattern classification, which has been well characterized when the number of features used for prediction is small. Diagonal LDA as its special case has recently been shown to be comparable to more complicated classifiers for microarray data [9,18]. In this method, the important parameters include prior probability, the mean vector of each class, and common covariance. They need to be estimated using training data. In this paper, a sparse diagonal discriminant analysis (SDDA) method is proposed, in which a maximum penalized likelihood estimation is employed by adding an L_1 penalty term to the likelihood function, resulting in a soft-thresholding on the mean vector, which realizes automatic biomarker identification. The penalized likelihood approach by an L_1 penalty is shown to be effective for variable selection [19,20]. On the breast cancer data set, we can identify some key genes (such as *ANXA1*) that were verified as correlating with breast cancer development and progression in biochemical or biomedical research and classify the breast cancer type correctly. More importantly, this model can give a good biological interpretation for gene selection in the classification step.

In our method, biomarker identification is an integral part of the classifier design algorithm. The identified biomarkers can directly optimize the performance of the final classifier. Microarray data contain many genes, but only a few dozen genes are informative; thus, direct application of the proposed method will incur prohibitive computational cost. To deal with this problem, we first take a preselected gene as a biomarker candidate based on the criterion of the maximal ratio of their between-groups to within-groups sum of squares (BSS/WSS). The BSS/WSS criterion is a useful tool to select genes and tends to pick up genes that are highly correlated.

The rest of this paper is organized as follows. In Section 2, we introduce our sparse diagonal discriminant analysis model, show the algorithm to estimate the parameters and give a mathematical explanation about biomarkers identification. In Section 3, we present the experimental results of the proposed method about biomarkers identification and cancer classification on breast cancers data. Finally, the discussion and future prospects are presented in Section 4.

2. Methods

We arrange the normalized gene expression data in a $p \times n$ matrix $\mathbf{X}$, where the entry x_{ij} represents the expression level of the ith gene in the jth sample, p is the number of genes and n is the number of samples, e.g., cancer cases. Suppose that all the columns in $\mathbf{X}$ arise from K classes $\{\omega_1, \omega_2, \ldots, \omega_K\}$ with the samples $\{\{x_{k1}, x_{k2}, \ldots, x_{kn_k}\}_{1 \leqslant k \leqslant K}\}$.

2.1. The diagonal linear discriminant analysis

The diagonal linear discriminant analysis (DLDA) method is a method based on componentwise normal assumption. The discriminant function is found by maximizing the class posteriors for optimal classification in the Bayes decision theory. Suppose that the probability of a sample x belonging to class ω_k is $P(\omega_k|x)$. Let the prior probability $P(\omega_k)$ of the kth class be π_k, where $0 \leqslant \pi_k \leqslant 1$, $\sum_{k=1}^{K} \pi_k = 1$. Let $P(x)$ and $P(x|\omega_k)$ be, respectively, the probability of sample x and the conditional probability.

We suppose that $P(x|\omega_k)$ satisfies the componentwise normal distribution with common variances:

$$P(x|\omega_k) = \frac{1}{(2\pi)^{p/2}\prod_{j=1}^{p}\sigma_j} \exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{(x^j-\mu_k^j)^2}{\sigma_j^2}\right\} \quad (1)$$

where μ_k^j is the jth component of the mean vector of the kth class ω_k, and σ_j^2 ($\sigma_j > 0$) is the variance of the jth component for all samples. Classification is achieved by assigning a pattern x to class ω_k for which the posterior probability $P(\omega_k|x)$ or the equivalent $\log(P(\omega_k|x))$ is the greatest.

By the Bayesian theorem, we have

$$P(\omega_k|x) = P(\omega_k)P(x|\omega_k)/P(x) \quad (2)$$

$$\begin{aligned}\log(P(\omega_k|x)) &= \log(P(\omega_k)) + \log(P(x|\omega_k)) - \log(P(x)) \\ &= -\frac{1}{2}\sum_{j=1}^{p}\frac{(x^j-\mu_k^j)^2}{\sigma_j^2} - \sum_{j=1}^{p}\log(\sigma_j) \\ &\quad + \log(\pi_k) - \left(\frac{p}{2}\log(2\pi) + \log(P(x))\right) \end{aligned} \quad (3)$$

The probability density function $P(x)$ is independent of class and is not required in the decision process. So, only the parameter $\theta = \{\pi_1, \ldots, \pi_{K-1},\ \{\mu_k^j\} 1 \leqslant k \leqslant K, 1 \leqslant j \leqslant p,\ \sigma_1, \ldots, \sigma_p\}$ needs to be estimated. Generally, θ is estimated by classical maximum likelihood estimation which maximizes the log-likelihood function for samples with known class labels.

2.2. Sparse diagonal discriminant analysis

From (3), we find that the label prediction of a probe depends only on the parameter θ, i.e., the complexity of DLDA is equivalent to the degree of freedom of $\theta : df(\theta) = Kp + p + K - 1$. For gene expression data, p is usually much larger than n, which means that the model has high complexity and, therefore, high prediction error when applied to a few training samples. Moreover, it is believed that only a few genes make discriminant contributions to cancer classification and the others are irrelevant or redundant, that is, the useful genes might be sparse.

In this paper, we propose to use a sparse diagonal discriminant analysis (SDDA) to automatically select the useful genes for classification with low complexity. The rationale is to impose an L_1 norm penalty on μ_k^j in the log-likelihood function:

$$\begin{aligned} l_p(\theta) &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\log(P(\omega_k|x_i)) - \lambda\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_k^j| \\ &= -\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\frac{(x_{ki}^j-\mu_k^j)^2}{\sigma_j^2} - n\sum_{j=1}^{p}\log(\sigma_j) \\ &\quad + \sum_{k=1}^{K} n_k\log(\pi_k) + C - \lambda\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_k^j| \end{aligned} \quad (4)$$

where $C = -\left(\frac{np}{2}\log(2\pi) + n\log(P(x))\right)$ is a constant, λ is a regularization parameter. It is easy to check whether the negative penalty item is proportional to the log density of the prior distribution for the parameter μ_k^j and the penalized log-likelihood is proportional to the log-posterior density. Our aim is to automatically shrink small estimates of μ_k^j to 0; thus, realizing biomarker identification, since the L_1-penalty yields a soft-thresholding rule with the sparsity property [19,20].

2.3. Solving SDDA by the EM algorithm

We explore the EM algorithm to calculate the parameter θ by maximizing the penalized log-likelihood function (4). Suppose that $\{\{x_{k1}, x_{k2}, \ldots, x_{kn_k}\}_{1\leqslant k\leqslant K}\}$ are the n training samples, and $x_{n+1}, \ldots, x_{n+m}$ are the m testing samples. Our EM algorithm begins by initializing a sample mean, sample variance and the prior probability as defined in (5)–(7), then it iterates between calculating their posterior probability $P(\omega_k|x_i)$ as in (8) and estimating the mean $\hat{\mu}_k^j$, variance $\hat{\sigma}_j^2$ and prior probability π_k as in (10)–(12) by maximizing $l_p(\theta)$, and finally ends with a mean estimation being convergent. The details are shown below:

(a) Initializing θ, for each $1 \leqslant k \leqslant K,\ 1 \leqslant j \leqslant p$,

$$^{(0)}\mu_k^j = \frac{\sum_{i=1}^{n_k} x_{ki}^j}{n_k} \quad (5)$$

$$^{(0)}\sigma_j^2 = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}^j - {}^{(0)}\mu_k^j)^2}{n} \quad (6)$$

$$^{(0)}\pi_k = \frac{n_k + m/K}{n+m} \quad (7)$$

(b) Updating the posterior probability of class assignment for the m probe samples, for each $n+1 \leqslant i \leqslant n+m,\ 1 \leqslant k \leqslant K$,

$$^{(t+1)}P(\omega_k|x_i) = \frac{{}^{(t)}\pi_k \times {}^{(t)}P(x_i|\omega_k, {}^{(t)}\mu_k^j, {}^{(t)}\sigma_j^2)}{\sum_{k'=1}^{K} {}^{(t)}\pi_{k'} \times {}^{(t)}P(x_i|\omega_{k'}, {}^{(t)}\mu_{k'}^j, {}^{(t)}\sigma_j^2)} \quad (8)$$

where

$$\begin{aligned} {}^{(t)}P(x_i|\omega_k, {}^{(t)}\mu_k^j, {}^{(t)}\sigma_j^2) &= \frac{1}{(2\pi)^{p/2}\prod_{j=1}^{p} {}^{(t)}\sigma_j} \\ &\quad \times \exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{(x_i^j - {}^{(t)}\mu_k^j)^2}{{}^{(t)}\sigma_j^2}\right\} \end{aligned} \quad (9)$$

is the conditional probability in the tth iteration.

(c) Updating θ by setting partial derivatives $\frac{\partial l_p(\theta)}{\partial\theta} = 0$, for each $1 \leqslant k \leqslant K,\ 1 \leqslant j \leqslant p$,

$$^{(t+1)}\mu_k^j = \mathrm{sgn}(\bar{\mu}_k^j)\left(|\bar{\mu}_k^j| - \frac{\lambda}{n_k} {}^{(t)}\sigma_j^2\right)_+ \quad (10)$$

$$^{(t+1)}\sigma_j^2 = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}^j - {}^{(t+1)}\mu_k^j)^2}{n} \quad (11)$$

$$^{(t+1)}\pi_k = \frac{n_k + \sum_{i=n+1}^{n+m} {}^{(t)}P(\omega_k|x_i)}{n+m} \quad (12)$$

where $\bar{\mu}_k^j = \frac{\sum_{i=1}^{n_k} x_{ki}^j}{n_k}$ can be computed before the iteration process. For any f, we denote the function $(\cdot)_+$ by setting $(f)_+ = f$ if $f > 0$, and $(f)_+ = 0$ otherwise.

(d) Repeating steps (b) and (c) until $\sum_{k=1}^{K}\sum_{j=1}^{p}|^{(t+1)}\ \mu_k^j - {}^{(t)}\mu_k^j| < \epsilon$, ϵ is a given threshold.

Then each probe sample x_i $(i = n+1, \ldots, n+m)$ is assigned to class ω_k, if

$$^{(t_e+1)}P(\omega_k|x_i) > {}^{(t_e+1)}P(\omega_{k'}|x_i), \quad \text{for all } k' \neq k \tag{13}$$

where $t_e + 1$ is the final iteration.

2.4. Rationale of SDDA for biomarker identification

It is noted that in our method μ_k^j and σ_j are estimated from the training samples, and they are updated to produce the sparse estimations of μ_k: if $\lambda > n_k|\bar{\mu}_k^j|/{}^{(t)}\sigma_j^2$, then $^{(t+1)}\mu_k^j = 0$; otherwise, $^{(t+1)}\mu_k^j$ is obtained by shrinking $\bar{\mu}_k^j$ towards 0 by an amount $\lambda^{(t)}\sigma_j^2/n_k$. Specially, if $\lambda = 0$, the model is degenerated into DLDA.

To reveal the relations among λ, sparsity of μ_k and biomarker identification, we assume $Z_\mu^\lambda \triangleq \{j|^{(t_e)}\mu_k^j = 0,\ k = 1, \ldots, K\}$ with respect to a fixed λ, then consider Eq. (9) after the final iteration:

$$\begin{aligned} ^{(t_e+1)}P(x_i|\omega_k, {}^{(t_e)}\mu_k^j, {}^{(t_e)}\sigma_j^2) &= \frac{1}{(2\pi)^{p/2}\prod_{j=1}^{p}{}^{(t_e)}\sigma_j}\exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{(x_i^j - {}^{(t_e)}\mu_k^j)^2}{{}^{(t_e)}\sigma_j^2}\right\} \\ &= \left[\frac{1}{(2\pi)^{p/2}\prod_{j=1}^{p}{}^{(t_e)}\sigma_j}\exp\left\{-\frac{1}{2}\sum_{j\in Z_\mu^\lambda}\frac{(x_i^j)^2}{{}^{(t_e)}\sigma_j^2}\right\}\right] \\ &\quad \exp\left\{-\frac{1}{2}\sum_{j\notin Z_\mu^\lambda}\frac{(x_i^j - {}^{(t_e)}\mu_k^j)^2}{{}^{(t_e)}\sigma_j^2}\right\} \end{aligned} \tag{14}$$

The expression in the square brackets "[]" is the same for all K classes, so it can be removed from the numerator and denominator of Eq. (8), that is, the genes in Z_μ^λ make no contribution to classification and can be removed. In this way, we arrive at biomarker identification and cancer classification simultaneously. Moreover, suppose that p_0 is the size of Z_μ^λ, the complexity of the SDDA model can be reduced into $d^\lambda = Kp + p + K - 1 - p_0$, that is, the complexity depends on the choice of λ.

2.5. Choice of the regularization parameter λ

Since λ is a tuning parameter to monitor the sparsity of μ_k, its choice should have some relationship with μ_k. Suppose $\Lambda = \{c_j \triangleq \max_{1\leqslant k\leqslant K}\left(\frac{n_k|\bar{\mu}_k^j|}{{}^{(0)}\sigma_j^2}\right),\ 1 \leqslant j \leqslant p\}$. For a given λ, if $c_j \leqslant \lambda$, the component j is shrunk to zeros for all μ_k, that is, the jth gene is removed in the first iteration. We rearrange Λ in increasing order, and let $c_{j_1}, c_{j_2}, \ldots, c_{j_\tau}$ be the first τ smallest elements. To properly expedite the sparsity process, we set λ as $c_{j_1}, c_{j_2}, \ldots, c_{j_\tau}$, which means that there are, respectively, $1, 2, \ldots, \tau$ removed genes in the first iteration, then we adopt v-fold cross-validation to minimize any misclassification error: for a fixed v, we randomly split the training dataset into v groups $g_1, \ldots, g_v$, which are of roughly equal size. The optimal λ is chosen to minimize the cross-validated misclassification error:

$$CV(\lambda) = \sum_{i=1}^{v}\sum_{x\in g_i}(L(x) \neq \hat{L}^{(-i)}(x)) \tag{15}$$

where $L(x)$ is the true label of x and $\hat{L}^{(-i)}(x)$ is the estimated label of x based on the training data without the ith group by SDDA. In our experiments, we set τ as 10% of the number of genes and $v = 7$.

3. Results

In this section, we apply the proposed method to the published breast cancer dataset [3] for biomarker identification and cancer classification. The dataset has an expression profile of 3226 genes on 22 breast tumor samples consisting of seven BRCA1, eight BRCA2 and seven sporadic cases from 21 patients. Here, we apply our method to classify BRCA1 versus BRCA2 and sporadic cases as indicated by Zhou et al. and Lee et al. [7,21].

We first employ the BSS/WSS criterion, i.e., the ratio of the between-group to the within-group sum of squares, to select the biomarker candidates that are more active in the experiment. Then, we perform the proposed method on the 200 biomarker candidates to discriminate the breast cancer type and identify biomarkers simultaneously.

3.1. Biomarker candidates selection

A major characteristic of microarray data is the very large number of dimensions or genes, most of which exhibit similar expression levels across samples. This may result in two problems: many gene profiles make no contribution to cancer classification and the computation complexity increases exponentially with the number of features. We thus perform a preliminary selection of genes using the BSS/WSS criteria, which are defined as the ratio of the total between-groups variance to within-groups variance, i.e., for each gene j,

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mu_k^j - \mu^j)^2}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki}^j - \mu_k^j)^2} \tag{16}$$

where μ^j denotes the average expression level of the jth gene across all samples. Lee [18] and Dudoit [9] stated that preselecting 200 genes does not affect the capability of most of the classifiers.

Fig. 1 represents 22 breast cancer samples consisting of seven BRCA1 (red circles) and 15 others (green stars) using the top three principal components of the total 3226 genes, the preselected 200 genes and the top 20 genes selected, respectively (see Table 1). In the left panel of Fig. 1, it can be seen that none of the 3226 genes provide any classification information, and the distribution just looks uniform in each direction. The middle panel of Fig. 1 shows that the preselected 200 genes can clearly separate different cancers. The right panel of Fig. 1 also

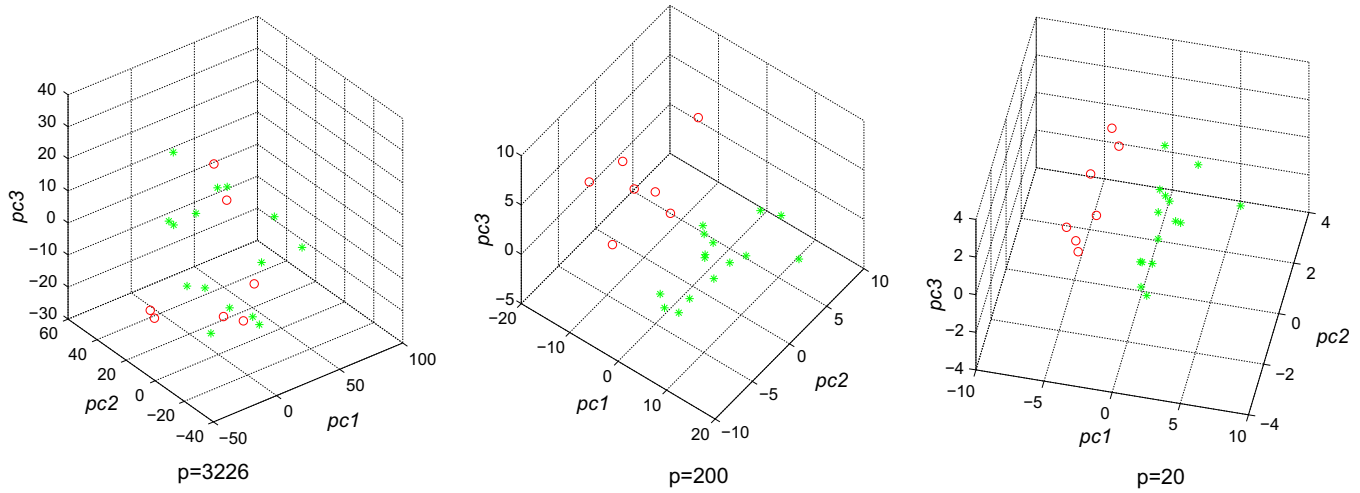

Fig. 1. Representation of 22 breast cancer samples consisting of seven BRCA1 (red circles) and 15 BRCA2 or sporadic cases (green stars). The left one uses the top three principal components of all the original 3226 genes; the middle one uses the top three principal components of the 200 preselected genes using the BSS/WSS criteria; and the right one uses the top three principal components of the 20 genes that were selected from Table 1.

shows the separability. It suggests that the 20 key genes we selected do contain the main classification information as the 200 preselected genes. It is worth noting that if the number of preselected genes is too small the real marker gene may be missed; if it is too large the computation complexity becomes high.

3.2. Biomarkers identification

Using the 200 preselected genes based on the BSS/WSS criterion, we employ SDDA to classify breast cancers and identify biomarkers. Table 1 describes the top 20 genes selected in 22 tests using our proposed method, and they are ranked by their selected frequency in column 1. We note that some leading genes in Table 1 have been verified as being related to breast cancer in biochemical or biomedical research. Annexin A1 (*ANXA1*) is the first gene in Table 1, whose expression at very low levels was observed in both cancerous and immortalized normal breast epithelial cell lines.

ANXA1 plays a significant regulatory role in human breast cancer development as its decreased or lost expression might occur at an early stage of malignant transformation [22]. The second gene in Table 1 is catenin (cadherin-associated protein) alpha 1 (102 kDa), which interacts with E-cadherin. The frequent alterations in alpha-catenin expression argue that loss of function in the E-cadherin-catenin pathway may be critical in the development of many breast cancers [23]. Ataxia-telangiectasia group D-associated protein (*ATDC*) is the fifth gene in Table 1. Recent research suggests that suppressed *ATDC* expression is associated with the malignant phenotype in tumor cell lines [24]. Some other strong genes (denoted by the 'MI' in column 3) are the same as the strong genes selected by the mutual information method [7]. Column 2 in Table 1 is the rank index of the BSS/WSS ratio [9]. Note that the valuable genes may not have a large ratio, such as *ANXA1* whose BSS/WSS rank index is 197, that is, the genes may

Table 1
Selected key genes.

Frequency	Rank of BSS/WSS index	ImageCloneID	Gene description
22	197	208718	Annexin A1 (*ANXA1*)
22	74	21652	Catenin (cadherin-associated protein), alpha 1 (102 kDa)
22	174	841093	Cullin 1
22	170	43021	Histidyl-tRNA synthetase (*HisRs*)
22	193	377275	Ataxia-telangiectasia group D-associated protein (*ATDC*)
21	121	823930	Actin-related protein 2/3 complex, subunit 1A (41 kDa)
21	76	240208	DKFZP434O125 protein
21	60	134748	Glycine cleavage system protein H (aminomethyl carrier)
21	178	815503	KIAA0071 protein
20	109	811108	Thyroid hormone receptor interactor 6
20	29	788721^{MI}	KIAA0090 protein
20	181	138604	MAD (mothers against decapentaplegic, *Drosophila*) homolog 2
20	114	471217	TIA1 cytotoxic granule-associated RNA-binding protein-like 1
18	7	26184	Phosphofructokinase, platelet
18	15	843076^{MI}	Signal transducing adaptor molecule (SH3 domain and ITAM motif) 1
18	52	45840	Splicing factor, arginine/serine-rich 4
17	42	47542	Small nuclear ribonucleoprotein D1 polypeptide (16 kDa)
17	11	897646^{MI}	Splicing factor, arginine/serine-rich 4
17	2	566887^{MI}	Chromobox homolog 3 (*Drosophila* HP1 gamma)
17	4	26082^{MI}	Very low density lipoprotein receptor

be missed if applying this criterion only to identify the top dozen key genes.

3.3. Visualization

Fig. 2 shows 22 breast cancer samples consisting of seven BRCA1 (red circles) and 15 others (green stars) using a number of key genes selected for their three-dimensional representation. Fig. 2(a) uses the leading three genes (*ANXA1*, *catenin-alpha1*, *cullin1*) from Table 1, while Fig. 2(b) uses the second to the fourth genes (*catenin-alpha1*, *cullin1*, *HisRs*) and Fig. 2(c) uses the third to the fifth genes (*cullin1*, *HisRs*, *ATDC*). These diagrams show that seven BRCA1 samples (red circles), 15 BRCA2 and sporadic (green stars) can be clearly separated using the three marker gene expression profiles. The genes selected may be useful as tumor markers and prognostic indicators. This result suggests that our proposed method is a useful strategy for discovering biomarkers with clinical relevance in cancer detection and disease classification.

3.4. Comparison of classification performance with other methods

Using the 200 preselected genes based on the BSS/WSS criterion, we evaluate the efficacy of our proposed method using the leave-one-out cross-validation method: first exclude a single sample for testing and predict the class label for the sample, then compare its prediction with the observed response.

In this paper, we compare the proposed method with Ye's RDA [25] and three simple but effective methods (DiagLDA, DiagQDA, KNN) as Dudoit et al. [9] and Lee et al. [18] used. As shown in the second row of Table 2, our method can predict their attribution successfully with no errors. However, under the same conditions, Ye's RDA and DiagQDA misclassify one sample out of 22 observations, and KNN has the worst performance with seven misclassified observations. Only DiagLDA does as the proposed method with no misclassification. However, DiagLDA does not perform biomarker identification while our method does. As shown in the third row of Table 2, the average number of features our method uses for classification is 54.59, while the other methods use a total of 200 features. This performance does verify the rationality of our method based on the diagonal discriminant analysis.

4. Discussion and conclusion

Biomarker identification and cancer classification are two important procedures in microarray gene expression data analysis. Some existing methods deal with them separately. There are at least two drawbacks. Firstly, the gene selection is independent of the resulting classifiers so that it is uncertain whether the selected subset of genes can lead to maximal prediction precision [17]. Secondly, though some feature extraction methods, such as LDA, produce the best features for classification, they do not select marker genes because the feature extracted is a combination of all genes. In this work, we have proposed a SDDA method, which maximizes penalized likelihood on diagonal LDA to overcome the drawbacks and to deal with biomarker identification and cancer classification in an integrated

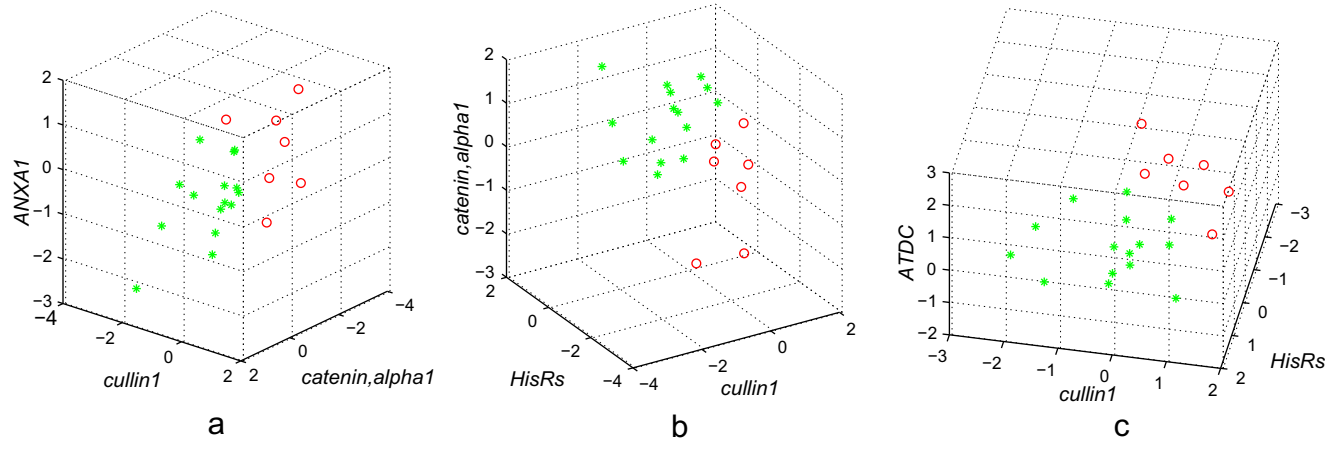

Fig. 2. Representation of 22 breast cancer samples consisting of seven BRCA1 (red circles) and 15 BRCA2 or sporadic cases (green stars). (a) uses the top three genes (*ANXA1*, *catenin-alpha1*, *cullin1*) from Table 2 for three-dimensional representation; (b) uses the second to the fourth genes (*catenin-alpha1*, *cullin1*, *HisRs*); and (c) uses the third to the fifth genes (*cullin1*, *HisRs*, *ATDC*).

Table 2
Classification performance.

Method	SDDA	DiagLDA	Ye's RDA	DiagQDA	KNN
Number of misclassified observations in the total 22 samples	0	0	1	1	7
Average number of features used for classification	54.59	200	200	200	200

framework. This method can provide good classification performance with automatic gene subset selection through the tuning of a penalty parameter.

Experimental results indicate that the SDDA method performs well in biomarkers identification and achieves high classification accuracy simultaneously. Moreover, this model has a mathematical interpretation of biomarkers identification in classification. Furthermore, visual inspection by marker genes is also made possible. Future work includes exploring the application of the L_1 penalized technique to other classifiers and other penalty terms such as 'elastic net' in Zou and Hastie [26]. We envisage that the integrated method will play an increasingly important role in the analysis of microarray data because of its superior performance in biomarker identification and cancer classification.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 60575004 and 10771220), the Ministry of Education of China (Grant No. SRFDP-20070558043) and the Hong Kong Research Grant Council (Project CityU 122607).

References

[1] Schmidt U, Begley CG. Cancer diagnosis and microarrays. Int J Biochem Cell Biol 2003;35:119–24.
[2] Marchionni L, Wilson RF, Wolff AC, et al. Systematic review: gene expression profiling assays in early-stage breast cancer. Ann Intern Med 2008;148(5):358–69.
[3] Hedenfalk I, Duggan D, Chen Y, et al. Gene expression profiles in hereditary breast cancer. N Engl J Med 2001;344:539–48.
[4] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
[5] Li L, Weinberg CR, Darden TA, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17:1131–42.
[6] Tibshirani R, Hastie T, Narasimhan B, et al. Class prediction by nearest shrunken centroids, with application to DNA microarrays. Stat Sci 2003;18:104–17.
[7] Zhou X, Wang X, Dougherty ER. Nonlinear probit gene classification using mutual information and wavelet-based feature selection. J Biol Syst 2004;12:371–86.
[8] Yang YY, Wang HY, Li X, et al. A feature ensemble technology to identify molecular mechanisms for distinction between multiple subtypes of lymphoma. Prog Nat Sci 2008;18:1491–500.
[9] Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.
[10] Desper R, Khan J, Schäffer AA. Tumor classification using phylogenetic methods on expression data. J Theor Biol 2004;228:477–96.
[11] Lin S. Mixture modeling of progression pathways of heterogeneous breast tumors. J Theor Biol 2007;249:254–61.
[12] Purohit PV, Rocke DM. Discriminant models for high throughput proteomics mass spectrometer data. Proteomics 2003;3:1699–703.
[13] Lilien RH, Farid H, Donald BR. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J Comput Biol 2003;10:925–46.
[14] Yang WH, Dai DQ, Yan H. Generalized discriminant analysis for tumor classification with gene expression data. In: Proceedings of the fifth international conference on machine learning and cybernetics (ICMLC 2006), Dalian, China; 2006. p. 4322–7.
[15] Yang WH, Dai DQ, Yan H. Feature extraction and uncorrelated discriminant analysis for high-dimensional data. IEEE Trans Knowledge Data Eng 2008;20:601–14.
[16] Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Brief Bioinform 2008;9:102–18.
[17] Li X, Rao SQ, Wang YD, et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Res 2004;32:2685–94.
[18] Lee JW, Lee JB, Park M, et al. An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal 2005;48:869–85.
[19] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 2001;96:1348–60.
[20] Pan W, Shen XT. Penalized model-based clustering with application to variable selection. J Mach Learn Res 2007;8:1145–64.
[21] Lee KE, Sha N, Dougherty ER, et al. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–7.
[22] Shen D, Chang HR, Chen Z, et al. Loss of annexin A1 expression in human breast cancer detected by multiple high-throughput analyses. Biochem Biophys Res Commun 2005;326:218–27.
[23] Pierceall WE, Woodard AS, Morrow JS, et al. Frequent alterations in E-cadherin and alpha- and beta-catenin expression in human breast cancer cell lines. Oncogene 1995;11:1319–26.
[24] Hosoi Y, Kappb LN, Murnane JP, et al. Suppression of anchorage-independent growth by expression of the ataxia-telangiectasia group D complementing gene, ATDC. Biochem Biophys Res Commun 2006;348:728–34.
[25] Ye JP, Wang T. Regularized discriminant analysis for high dimensional, low sample size data. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2006); 2006. p. 454–63.
[26] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005;67:301–20.