BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 15 2006, pages 1855–1862
doi:10.1093/bioinformatics/btl190

Gene expression

Independent component analysis-based penalized discriminant method for tumor classification using gene expression data

De-Shuang Huang[1,*] and Chun-Hou Zheng[1,2]

[1]Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, PO Box 1130, Hefei, Anhui 230031, China and [2]Department of Automation, University of Science and Technology of China, China

Received on November 21, 2005; revised on April 27, 2006; accepted on May 11, 2006
Advance Access publication May 18, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Microarrays are capable of determining the expression levels of thousands of genes simultaneously. One important application of gene expression data is classification of samples into categories. In combination with classification methods, this technology can be useful to support clinical management decisions for individual patients, e.g. in oncology. Standard statistic methodologies in classification or prediction do not work well when the number of variables *p* (genes) far too exceeds the number of samples *n*. So, modification of existing statistical methodologies or development of new methodologies is needed for the analysis of microarray data.

Results: This paper proposes a new method for tumor classification using gene expression data. In this method, we first employ independent component analysis to model the gene expression data, then apply optimal scoring algorithm to classify them. Further speaking, this approach can first make full use of the high-order statistical information contained in the gene expression data. Second, this approach also employs regularized regression models to handle the situation of large numbers of correlated predictor variables. Finally, the predictive models are developed for classifying tumors based on the entire gene expression profile. To show the validity of the proposed method, we apply it to classify four DNA microarray datasets involving various human normal and tumor tissue samples. The experimental results show that the method is efficient and feasible.

Availability: Matlab scripts are available on request.

Contact: dshuang@iim.ac.cn

INTRODUCTION

A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. Current methods for classifying human malignancies are mostly to rely on a variety of morphological, clinical and molecular variables. Despite recent progress, there are still many uncertainties in diagnosis. Furthermore, it is likely that the existing classes of the tumors are heterogeneous and comprise diseases that are molecularly distant. Recently, with the development of large-scale high-throughput gene expression technology, it has become possible for ones to diagnose and classify diseases, particularly cancers, directly based on these DNA microarray technologies (Alizadeh *et al*., 2000). This technique has been termed as 'class prediction' in the microarray literature (Golub *et al*., 1999). By monitoring the expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of the molecular variations among tumors, and hence to a finer and more reliable classification.

With the wealth of gene expression data from microarrays being produced, more and more new prediction, classification and clustering techniques are being used for analysis of the data. Up to now, several studies have been reported on the application of microarray gene expression data analysis for molecular classification of cancer (Alon *et al*., 1999; Bittner *et al*., 2000; Furey *et al*., 2000). And, the analysis of differential gene expression data has been used to distinguish between different subtypes of lung adenocarcinoma (Bhattacharjee *et al*., 2001) and colorectal neoplasm (Selaru *et al*., 2002). Also, the work that predicts clinical outcomes in breast cancer (van't Veer *et al*., 2002; West *et al*., 2001) and lymphoma (Shipp *et al*., 2002) from gene expression data has been proven to be successful. Golub *et al*. (1999) utilized a nearest-neighbor classifier method for the classification of acute myeloid lymphoma (AML) and acute leukemia lymphoma (ALL) in children. Dudoit *et al*. (2002) performed a systematic comparison of several discrimination methods for classification of tumors based on microarray experiments. While linear discriminant analysis was found to perform the best, in order to utilize the method, the number of genes selected had to be drastically reduced from thousands to tens using a univariate filtering criterion.

One feature of microarray data is that the number of tumor samples collected tends to be much smaller than the number of genes. The number for the former tends to be on the order of tens or hundreds, while microarray data typically contain thousands of genes on each chip. In statistical terms, it is called 'large *p*, small *n*' problem (West, 2003), i.e. the number of predictor variables is much larger than the number of samples. In theory, the more recent technique, support vector machines (SVM), should be more suitable for this problem. Furthermore, Furey *et al*. (2000) have applied SVM to classify tumors using microarray data. In fact, although SVM has been successfully applied to some other problems, it requires more training than the linear discriminant analysis. Also, the generalization of the SVM to classify more than two classes of problems is not solved significantly.

Ghosh (2003) proposed a methodology using regularized regression models for the classification of tumors. In this literature, he focused on three types of regularized regression models, i.e. ridge regression, principal components regression and partial least

*To whom correspondence should be addressed.

squares regression. One drawback of these techniques is that only second-order statistical information of the gene data is used. However, in the task such as classification, much of the important information may be contained in the high-order relationships among samples. And thus, it is important to investigate whether or not the generalizations of principal component analysis (PCA), which are sensitive to high-order relationships (not just second-order relationships), are advantageous. Usually, ICA (Bartlett, *et al*., 2002; Teschendorff *et al*., 2005) is one of such generalizations. A number of algorithms for performing ICA have been proposed. Please see literature (Hyvärinen *et al*., 2001) for the details of these techniques. Here, we shall employ FastICA, which was proposed by Hyvärinen (1999) and proven successful in many applications, to address the problems of tumor classification.

In this article, we present a new methodology that combines ICA and regularized regression models (Frank and Friedman, 1993) for analyzing gene expression data. We first perform ICA on gene expression data, then apply optimal scoring algorithm (Hastie *et al*., 1994) to classify the gene expression data. The advantages of this approach are that first, we can make full use of the high-order statistical information contained in the gene expression data; second, regularized regression models can handle the situation of large numbers of correlated predictor variables; finally, we can develop predictive models for classifying tumors based on the entire gene expression profile. To validate the efficiency, the proposed method is applied to classify four different DNA microarray datasets including colon cancer data (Alon *et al*., 1999), acute leukemia data (Golub *et al*., 1999), hepatocellular carcinoma data (Iizuka *et al*., 2003) and high-grade glioma data (Nutt *et al*., 2003). The prediction results show that our method is efficient and feasible.

METHODS

Independent component analysis

ICA is a useful extension of PCA that has been developed in context with blind separation of independent sources from their linear mixtures (Comon, 1994). Such blind separation techniques have been used, e.g. in various applications of auditory signal separating, medical signal processing and so on. Roughly speaking, rather than requiring that the coefficients of a linear expansion of the data vectors be uncorrelated as in PCA, in ICA these coefficients must be mutually independent (or as independent as possible). This implies that higher-order statistics are needed in determining the ICA expansion.

Considering an $n \times p$ data matrix $\mathbf{X}$, whose rows $\mathbf{r}_i$ (i=1,$\cdots$,n) correspond to observational variables and whose columns $\mathbf{c}_j$ (j=1,$\cdots$,p) are the individuals of the corresponding variables, the ICA model of $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

Without loss of generality, $\mathbf{A}$ is an $n \times n$ mixing matrix and $\mathbf{S}$ is an $n \times p$ source matrix subject to the condition that the rows of $\mathbf{S}$ are as statistically independent as possible. Those new variables contained in the rows of $\mathbf{S}$ are called as 'independent components', i.e. the observational variables are linear mixtures of independent components. The statistical independence between variables can be quantified by mutual information $I = \sum_k H(s_k) - H(S)$, where $H(_{sk})$ is the marginal entropy of the variable s_k and $H(S)$ is the joint entropy. Estimating the independent components can be accomplished by finding the right linear combinations of the observational variables, since we can invert the mixing matrix as

$$\mathbf{U} = \mathbf{S} = \mathbf{A}^{-1}\mathbf{X} = \mathbf{WX}. \tag{2}$$

There are a number of algorithms for performing ICA (Comon, 1994; Hyvärinen, 1999; Zheng, *et al*., 2005, 2006). In this paper, we shall employ the FastICA algorithm, which was proposed by Hyvärinen (1999), to address the problems of tumor classification. In this algorithm, the mutual information is approximated by a 'contrast function':

$$J(s_k) = (E\{G(s_k)\} - E\{G(v)\})^2 \tag{3}$$

where G is an arbitrary non-quadratic function and v is a normally distributed variable. The interested readers can refer to literature (Hyvärinen, 1999) for further details.

Like PCA, ICA can remove all linear correlations. By introducing a non-orthogonal basis, it also takes into account higher-order dependencies in the data. Particularly, ICA is in a sense superior to PCA, which is just sensitive to second-order relationships of the data. And, the ICA model usually leaves some freedom of scaling and sorting by convention, the independent components are generally scaled to unit deviation, while their signs and orders can be chosen arbitrarily.

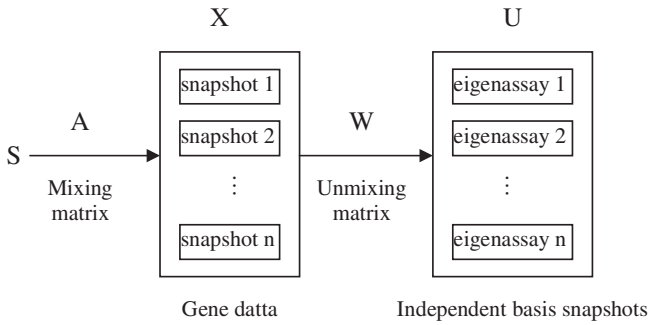

Fig. 1. The first gene expression data synthesis model. To find a set of independent basis snapshots (eigenassay), the snapshots in $\mathbf{X}$ are considered to be a linear combination of statistically independent basis snapshots (eigenassay, the rows in $\mathbf{S}$), where $\mathbf{W}$ is the unmixing matrix and $\mathbf{A} = \mathbf{W}^{-1}$ is an unknown mixing matrix. The independent eigenassay is estimated as the output $\mathbf{U}$ of the learned ICA.

ICA models of gene expression data

Now let the $n \times p$ matrix $\mathbf{X}$ denote the gene expression data (generally speaking, $n \ll p$), x_{ij} is the expression level of the j-th gene in the i-th assay. $\mathbf{r}_i$ (a p-dimensional vector), the i-th row of $\mathbf{X}$, denotes the snapshot of the i-th assay (cell sample) (In the gene data literature, the problem is usually formulated using the transposed matrix $\mathbf{X}^{\mathrm{T}}$). Alternatively, $\mathbf{c}_j$ (an n-dimensional vector), the j-th column of $\mathbf{X}$, is the expression profile of the j-th gene. We suppose that the data have already been preprocessed and normalized, i.e. every sample has mean zero and standard deviation one.

Regardless of which algorithm is used to compute ICA, we can apply ICA to model gene expression data as shown in Figure 1. In this model, the snapshots $\mathbf{r}_i$ in $\mathbf{X}$ are considered to be a linear mixture of statistically independent basis snapshots (eigenassay) $\mathbf{S}$ combined by an unknown mixing matrix $\mathbf{A}$. The ICA algorithm learns the weight matrix $\mathbf{W}$, which is used to recover a set of independent eigenassays in the rows of $\mathbf{U}$. In this architecture, the snapshots $\mathbf{r}_i$ are variables and the gene expression profile values provide observations for the variables. Essentially, this method coincides with the traditional ICA-like model of cock-tail problem (Comon, 1994). Projecting the input snapshots onto the learned weight vectors produces the independent basis snapshots. As a result, the corresponding mixing and unmixing models can be represented as follows:

$$\mathbf{X} = \mathbf{AS} \tag{4}$$

$$\mathbf{U} = \mathbf{S} = \mathbf{A}^{-1}\mathbf{X} = \mathbf{WX} \tag{5}$$

Fig. 2. The second gene expression data synthesis model. Each gene profile in the data matrix was considered to be a linear combination of underlying basis expression profiles (eigengenes) in the matrix **A** (the columns in **A**). Each of the basis expression profiles was associated with a set of independent 'causes (coefficient)', given by a vector of coefficients in S. The basis profiles were estimated by $\mathbf{A} = \mathbf{W}^{-1}$, where **W** is the learned ICA weight matrix.

In this approach, ICA is used to find a matrix **W** such that the rows of **U** are as statistically independent as possible. The independent eigenassays estimated by the rows of **U** are then used to represent the snapshots. The representation of the snapshots consists of their corresponding coordinates with respect to the eigenassays defined by the rows of **U**, i.e.

$$\mathbf{r}_j = a_{j1}\mathbf{u}_1 + a_{j2}\mathbf{u}_2 + \cdots + a_{jn}\mathbf{u}_n \tag{6}$$

These coordinates are contained in the rows of mixing matrix $\mathbf{A}=\mathbf{W}^{-1}$. Clearly, every coordinate $\mathbf{a}_j$ (row of **A**) is an n-dimensional vector while the snapshot $\mathbf{r}_j$ is a p-dimensional vector. In general, the number of genes in a single assay is in the thousands while the number of assay is up to hundreds. So the above procedure can be used to compress the gene expression data.

From another viewpoint, the gene expression profiles (columns of **X**) can be regarded as points in a multidimensional space with dimensions corresponding to the number of samples. The linear ICA model $\mathbf{X}=\mathbf{AS}$ represents the gene expression profiles (the columns of **X**) by a new set of basis vectors (the columns of **A**, Fig. 2). This idea is based on the assumptions that, first, the gene expression profiles are determined by a combination of hidden regulatory variables, which were called 'expression modes'. Second, the genes' responses to these variables can be approximated by linear functions (Liebermeister, 2002; Hori *et al*., 2001). Expression mode k is characterized by its profile over the samples (k-th column of **A**) and by its linear influences on the genes (k-th row of **S**). In this paper, we just use this idea to find a good set of basis profiles (eigengenes) to represent gene expression data so that they can be reasonably regularized.

Search for the consensus eigenassays

Chiappetta (2004) has pointed out that unlike PCA, ICA requires searching for the maxima of a target function in a large-dimensional configuration space. Therefore, one often encounters difficulties with local maxima in which most algorithms may get stuck, and the result may be sensitive to initialization. We also find in the experiments that compared with PCA, ICA is not always reproducible when used to analyze gene expression data. This problem had also been found by Liebermeister (2002). In addition, the results obtained from an ICA algorithm are not 'ordered'. In the literature (Chiappetta *et al*., 2004), the authors had considered that, the reason of this phenomenon is that the ICA algorithm may converge to local optima. In addition, they have given out a 'consensus source (eigenassay)' search algorithm which yields extremely stable and robust estimates for the eigenassays, as well as indications relative to their stability.

In this paper, we use the method advised by Chiappetta *et al*. (2004) to overcome these difficulties, which uses the following procedure. The independent source estimation is run several times (say, 100 times), with different random initializations, and 'consensus sources' are recorded namely, eigenassays which are obtained with a frequency larger than a certain threshold are conserved, and their frequencies of appearance are recorded and used as 'credibility indices.' As a result, one is led to a (variable, data-driven) number of average consensus eigenassays $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

Finally, the corresponding consensus mixing matrix **A** is computed as follows:

$$a_{ji} = \sum_{k=1}^{n} v_{ik}(\bar{\mathbf{s}}_i)' r_j, \quad j = 1, \ldots, n,\ i = 1, \ldots, n. \tag{7}$$

Here **V** is the inverse of the $n \times n$ matrix **C** of the scalar product of the consensus eigenassays($c_{ij} = (\bar{\mathbf{s}}_i)'\bar{\mathbf{s}}_j$). For more details, please see the literature (Chiappetta *et al*., 2004).

Interpretation of ICA results

The ICA model states that different modes exert independent influences on the genes. To interpret in more detail, the first step of the analysis is the study of the mixing matrix **A**. For a fixed eigenassay, say eigenassay i, the coefficients a_{ji} represent the projection of snapshot j on source i, or the 'importance' of eigenassay i in snapshot j. If one believes in the 'linear mixture of independent eigenassay' model and accepts identifying a source with a regulation pathway in first approximation, the coefficients a_{ji} would allow one to assert to which extent the eigenassay i was (positively or negatively) 'active' in snapshot j.

In addition, the distribution of the values of the column of the mixing matrix **A** is often interesting and may reveal specific features of the dataset. Particularly interesting is the situation where the distribution of mixing coefficients for a given eigenassay exhibit a bimodal or multimodal behavior. This indicates that the source under consideration has a good discriminating power between two or more different classes of conditions. However, as Chiappetta (Chiappetta *et al*., 2004) has pointed out even though bimodal distributions yield spectacular results, good discrimination may also be obtained without such a behavior.

A second step in the interpretation of ICA results is to analyze carefully the behavior of specific genes in different eigenassays. It generally happens that a given independent eigenassay is characterized by a number of significantly overexpressed (or underexpressed) genes. Putting such genes into correspondence with snapshots, or clinical data, may happen to be extremely informative. Because the main aim of this paper is not to study biological interpretation of ICA results for microarray data, moreover, there have been many literatures concern this issue, hence we will not discuss it in detail here. Readers who are interested in this issue can refer literatures further (Liebermeister, 2002; Hori *et al*., 2001; Martoglio *et.al*., 2002; Chiappetta *et al*., 2004).

ICA and PCA

PCA can be derived as a special case of ICA, which uses Gaussian source models. The assumption of Gaussian sources implicit in PCA makes it inadequate when the true sources are non-Gaussian. In particular, we have empirically observed that many gene expression data are 'sparse' or 'super-Gaussian' signals (all the four datasets used in this paper are 'super-Gaussian' signals). When sparse source models are appropriate, ICA has the following potential advantages over PCA: (1) It provides a better probabilistic model of the data, which better identifies where the data concentrate in n-dimensional space. (2) It uniquely identifies the mixing matrix **A**. (3) It finds an unnecessarily orthogonal basis which may reconstruct the data better than PCA in the presence of noise. (4) It is sensitive to high-order statistics in the data, not just the covariance matrix (Bartlett *et al*., 2002).

Figure 3 illustrates these points with an example. The figure shows samples from a three-dimensional (3D) distribution constructed by linearly mixing two high-kurtosis sources. The figure shows the basis vectors found by PCA and by ICA on this problem. Since the three ICA basis vectors are non-orthogonal, they change the relative distance between data points. This change in metric may be potentially useful for classification algorithms, like nearest neighbor, that make decisions based on relative distances between points. The ICA basis also alters the angles between data points, which affects similarity measures such as cosines. Moreover, if an undercomplete basis set is chosen, PCA and ICA may span different subspaces.

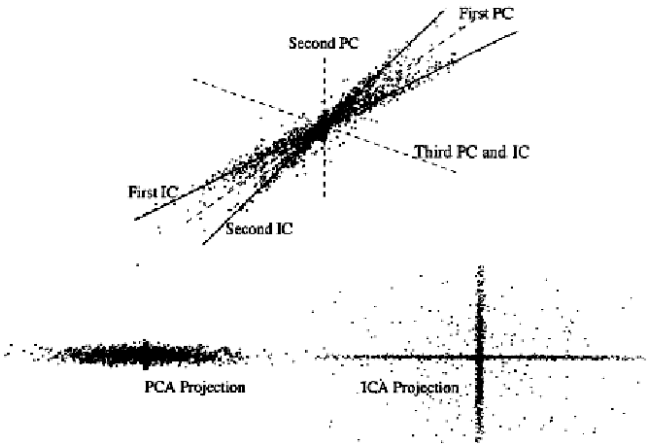

Fig. 3. (top) Example 3D data distribution and corresponding PC and IC axes. Each axis is a column of the mixing matrix **A** found by PCA or ICA. Note the PC axes are orthogonal while the IC axes are not. If only two components are allowed, ICA chooses a different subspace than PCA. (bottom left) Distribution of the first PCA coordinates of the data. (bottom right) Distribution of the first ICA coordinates of the data. Note that since the ICA axes are non-orthogonal, relative distances between points are different in PCA than in ICA, as are the angles between points (Bartlett *et al*., 2002).

For example, in Figure 3, when only two dimensions are selected, PCA and ICA choose different subspaces (Bartlett *et al*., 2002).

It should be noted that ICA is a very general technique. When super-Gaussian sources are used, ICA can be seen as doing something akin to non-orthogonal PCA and to cluster analysis, however, when the source models are sub-Gaussian, the relationship between these techniques is less clear. See (Lee *et al*., 1999) for a discussion of ICA in the context of sub-Gaussian sources.

Penalized regression models

In this section, we briefly outline two types of regularized regression models, i.e. ridge regression and principal component regression.

Ridge regression Consider the standard regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{8}$$

where $\mathbf{y}$ is an n-dimensional response vector; $\mathbf{X}$ is an $n \times p$ predictor matrix; β is a p-dimensional vector of unknown regression parameters; ε is a random vector with zero mean and one variance. In this paper, because n is smaller than p, the usual ordinary least squares estimator will not be well defined. An alternative is to use the ridge regression estimator of β in Equation (8):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{9}$$

where $\mathbf{I}$ is a $p \times p$ identity matrix and λ is a constant. The parameter λ controls the amount of shrinkage in the data.

Principal component regression The method of principal components regression can be traced back to the literature (Massy, 1965). To use this method, we first perform a singular value decomposition of the gene data $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} \tag{10}$$

where $\mathbf{U}$ is the $n \times n$ singular value decomposition matrix, and it has both orthonormal rows and columns. The diagonal matrix $\mathbf{D}$ contains the ordered eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ on the diagonal elements so that $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, where $d_1 > d_2 > \cdots > d_n > 0$. $\mathbf{V}$ is a $p \times n$ matrix matrix with orthonormal columns. Plugging this decomposition into Equation (8), we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{H}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{11}$$

where H=UD and $\gamma = \mathbf{V}^{\mathrm{T}}\beta$. We can fit the model in Equation (11) using ordinary least squares and get an estimate of β by multiplying $\mathbf{V}$ to the least squares estimator of γ in Equation (11).

Optimal scoring

In the previous section, we have described two penalized regression models that have been used successfully in other applications such as in chemometrics. However, in this paper, the goal for our interest is classification. Thus, we should firstly re-express the classification problem as a regression problem. This is done using the optimal scoring algorithm. The point of optimal scoring is to turn categorical variables into quantitative ones by assigning scores to classes (categories).

Let g_i denote the tumor class for the i-th sample (i=1, . . . , n), we assume that there are G tumor classes so that g_i takes values { 1, . . . , G }. We first convert $\mathbf{g} = [g_1, \ldots, g_n]^{\mathrm{T}}$ into an $n \times G$ matrix $Y = [Y_{ij}]$, where $Y_{ij} = 1$ if the i-th sample falls into class j, and 0 otherwise. Let $\theta_k(\mathbf{g}) = [\theta_k(g_1), \cdots \theta_k(g_n)]^T$ ($k = 1, \ldots, G$) be the $n \times 1$ vector of quantitative scores assigned to $\mathbf{g}$ for the k-th class. The optimal scoring problem involves finding the coefficients β_{k} and the scoring maps θ_{k} that minimize the following average squared residual:

$$\mathrm{ASR} = \frac{1}{n}\sum_{k=1}^{G}\sum_{i=1}^{n}(\theta_k(g_i) - \mathbf{X}_i^{\mathrm{T}}\beta_k)^2 \tag{12}$$

Let $\Theta_{G\times J}$ ($J \leq G-1$) be a matrix of j score vectors for the G classes, i.e. its k-th row is the scores, $\theta_k(\cdot)$. Assume that the minimization of Equation (12) is subject to the constraint$n^{-1}\|\mathbf{Y}\boldsymbol{\Theta}\|^2 = 1$, then, as mentioned by Hastie *et al*. (1994), the minimization of this constrained optimization problem leads to the estimates of β_{k} that are proportional to the discriminant variables in linear discriminant analysis. The interested readers can refer to the literatures (Hastie *et al*., 1994, 1995).

Penalized optimal scoring for classification

So far, we have outlined the components necessary for the implementation of our procedure. In this section, we give out our algorithm for classifying the tumor samples.

We propose to use ICA for regularizing the gene expression data and then use a penalized optimal scoring procedure for classification. The outline of our method is shown as follows:

Step 1: Using the ICA model $\mathbf{X} = \mathbf{A}\mathbf{S}$ to present the gene expression data, i.e. using ICA and consensus sources algorithm to calculate the eigengenes (columns of $\mathbf{A}$) and the independent coefficients (rows of $\mathbf{S}$).

Step 2: Choose an initial score matrix $\Theta_{G\times j}$ with $J \leq G-1$ satisfying $\Theta^{\mathrm{T}}D_p\Theta = I$, where $D_{\mathrm{p}} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}/n$. Let $\Theta_0 = \mathbf{Y}\Theta$.

Step 3: Fit a multivariate penalized regression model of Θ_0 on $\mathbf{A}$, yielding the fitted values $\hat{\boldsymbol{\Theta}}_0$ and the fitted regression function $\hat{\boldsymbol{\eta}}_0(A)$. Let $\hat{\boldsymbol{\eta}}(\mathbf{X}) = S^{+}\hat{\boldsymbol{\eta}}_0(\mathbf{A})$ be the vector of the fitted regression function on $\mathbf{X}$, where S^{+} is the pseudoinverse of $\mathbf{S}$.

Step 4: Obtain the eigenvector matrix Φ of $\Theta_0^{\mathrm{T}}\hat{\boldsymbol{\Theta}}_0$, and hence the optimal scores $\Theta_1 = \Theta\Phi$.

Step 5: Let $\eta(X) = \Phi^{\mathrm{T}}\hat{\boldsymbol{\eta}}(X)$.

What should be explained is that the objective function we are minimizing in *Step* 3 is the following expression:

$$\mathrm{ASR} = \frac{1}{n}\sum_{k=1}^{G}\sum_{i=1}^{n}(\theta_k(g_i) - \mathbf{A}_i^{\mathrm{T}}\gamma_k)^2 \tag{13}$$

where $\mathbf{A}_i$ is the i-th column of $\mathbf{A}$, $\gamma_k = \mathbf{S}\beta_{\mathrm{k}}$. Another problem is how to choose the initial values for Θ. Readers can refer to the discussions about this problem in literature (Hastie *et al*., 1994). In fact, our algorithm is somewhat similar to the algorithm proposed by Ghosh (2003), except that we replace principal components with independent components.

Once the algorithm has been run, we now have a discriminant rule for classifying new samples. We use the nearest centroid rule to form the

classifier, i.e. assign a new sample $\mathbf{X}_{\text{new}}$ to the class j that minimizes

$$\delta(\mathbf{X}_{\text{new}}, j) = \|\mathbf{D}(\eta(\mathbf{X}_{\text{new}}) - \bar{\boldsymbol{\eta}}^j\|^2 \quad (14)$$

where $\bar{\boldsymbol{\eta}}^j = [\sum \eta(\mathbf{X}_i)]/n_j$ denotes the fitted centroid of the j-th class. $\mathbf{D}$ is a matrix with diagonal element

$$D_{kk} = \left(\frac{1}{\lambda_k^2(1-\lambda_k^2)}\right)^{1/2}, \quad (15)$$

where λ_k is the k-th largest eigenvalue calculated in *Step* 4 of the algorithm.

Choosing the optimal amount of regularization

Ridge regression, principal components regression and independent component regression models involve a regularization parameter that must be selected in advance. In ridge regression method, the regularization parameter is λ, while for principal components, the parameter that needs to be chosen is the number of components included in the model. For ICA regression model, the parameters that need to be chosen are both the number and the subset of eigengenes (columns of $\mathbf{A}$) included in the model. In contrast to PCA method, where feature (principal component) subset selection is based on energy criterion, the selection of an ICA basis subset is not immediately obvious since the energies of the independents cannot be determined. Furthermore, it is conjectured that some feature selection scheme focused on 'recognition' rather than on 'reconstruction' could augment the classification performance. With this goal in mind, we used the sequential floating forward selection (SFFS) technique (Feri *et al*., 1994) to find the most discriminating ICA features (columns of $\mathbf{A}$, every eigengene corresponding an engenassay).

For this SFFS method, features are selected successively by adding the locally best feature points, which provide the highest incremental discriminatory information, to the exiting feature subset. In addition, the SFFS method goes through cleaning periods, in which features are removed systematically so long as the performance is improved after pruning. We use leave-one-out cross-validation in the training dataset to determine the number of components to include in the model. Readers who want to know the details about SFFS can refer to literature (Feri *et al*., 1994).

RESULTS

In this section, we shall demonstrate the efficiency and effectiveness of the proposed methodology described above by classifying four datasets with various human tumor samples.

Datasets

In this study, four publicly available microarray datasets are used to study the tumor classification problem. They are colon cancer data (Alon *et al*., 1999), acute leukemia data (Golub *et al*., 1999), hepatocellular carcinoma data (Iizuka *et al*., 2003) and high-grade glioma data (Nutt *et al*., 2003), respectively. In these datasets, all data samples have already been assigned to a training set or test set.

An overview of the characteristics of all the datasets can be found in Table 1. The acute leukemia data in literature (Golub *et al*., 1999) have already been used frequently in previous microarray data analysis studies. Preprocessing of this dataset was done by setting threshold and log-transforming on the original data, similar to the one introduced in the original publication. Threshold technique is generally achieved by restricting gene expression levels to be larger than 20. In other words, the expression levels which are smaller than 20 will be set to 20. Regarding the log-transformation, the natural logarithm of the expression levels usually is taken. In addition, no further preprocessing is applied to the rest of the datasets.

Table 1. Summary of the datasets for the four binary cancer classification problems

D	TR		TE		Levels	M
	C1	C2	C1	C2		
1	14	26	8	14	2000	T1
2	11	27	14	209	7129	T1
3	12	21	8	19	7129	T1
4	21	14	14	15	12 625	T1

D, datasets; TR, training set; TE, test set; C1, class 1; C2, class 2; levels, the number of genes; M, microarray technology; T1, oligonucleotide; 1, colon cancer data; 2, acute leukemia data; 3 hepatocellular carcinoma data; 4, high-grade glioma data.

Experimental results

We now use the proposed methodology to classify the tumor data. Since all data samples in these four datasets have already been assigned to a training set or test set, we built the classification models using the training samples and estimated the classification correct rates using the test set.

To obtain reliable experimental results showing comparability and repeatability for different numerical experiments, this study not only uses the original division of each dataset in training and test set, but also reshuffles all datasets randomly. In other words, all numerical experiments were performed with 20 random splitting of the four original datasets. And, they are also stratified, which means that each randomized training and test set contains the same amount of samples of each class compared with the original training and test set.

We used penalized independent component regression (P-ICR) proposed in this paper to analyze the four gene expression datasets. In the experiment, we have sought as many independent components as tumor samples in every run of ICA and as many consensus eigenassays as tumor samples. Also, before choosing the eigenassays with SFFS, the eigenassays with credibility <20% are deleted. For comparison, we also used penalized ridge regression (P-RR), penalized principal component regression (P-PCR) proposed in (Ghosh, 2003) and PAM (Tibshirani *et al*., 2002) to do the same tumor classification experiment.

The classification results for tumor and normal tissues using our proposed penalized methods are displayed in Table 2. For each classification problem, the experimental results gave the statistical means and standard deviations of accuracy on the original dataset and 20 randomizations as described above. Since the random splits for training and test set are disjoint, the results given in Table 2 are unbiased and can in general also be too optimistic.

To show the efficiency and feasibility of the method proposed in this paper, the results using other 9 methods (Methods 1–9) are also listed in Table 2 for comparison. These 9 methods can be subdivided in two steps: dimensionality reduction and classification. For dimensionality reduction, classical PCA as well as kernel PCA (with linear or RBF kernel) are used. Fisher discriminant analysis (FDA) and least squares support vector machine (LS-SVM) are then used for classification. Note that these methods and results were ever reported in literature (Pochet *et al*., 2004), where the divisional method of each training and test dataset is the same as ours. Readers can see the details about the first 9 methods from literature (Pochet *et al*., 2004).

Table 2. The summary of the results numerical experiments on four datasets

Experiments No.	Methods	Colon data Training set	Colon data Test set	Leukemia data Training set	Leukemia data Test set	Hepatocellular data Training set	Hepatocellular data Test set	Glioma data Training set	Glioma data Test set
1	LS-SVM linear kernel	99.64 ± 0.87	82.03 ± 7.49	100.00 ± 0.00	92.86 ± 4.12	73.88 ± 16.21	68.43 ± 4.52	90.02 ± 14.16	61.25 ± 11.75
2	LS-SVM RBF kernel	98.33 ± 2.36	81.39 ± 9.19	100.00 ± 0.00	93.56 ± 4.12	87.16 ± 16.73	68.61 ± 6.32	98.41 ± 7.10	69.95 ± 8.59
3	LS-SVM linear kernel (no regularization)	49.40 ± 8.93	51.73 ± 12.19	93.61 ± 15.93	87.39 ± 14.61	53.82 ± 5.68	49.56 ± 12.60	50.79 ± 12.75	48.93 ± 10.88
4	PCA + FDA (unsupervised PC selection)	90.95 ± 5.32	80.30 ± 9.65	99.50 ± 1.31	94.40 ± 3.84	89.61 ± 9.92	68.25 ± 7.37	92.29 ± 7.12	68.72 ± 7.24
5	PCA + FDA (supervised PC selection)	95.24 ± 5.56	76.84 ± 7.41	99.50 ± 1.31	93.56 ± 4.59	90.33 ± 11.52	66.67 ± 9.96	92.97 ± 10.14	65.52 ± 11.01
6	kPCA lin + FDA (unsupervised PC selection)	90.95 ± 5.32	80.30 ± 9.65	99.50 ± 1.31	94.40 ± 3.84	89.61 ± 9.92	68.25 ± 7.37	92.52 ± 6.98	68.31 ± 6.78
7	kPCA lin + FDA (supervised PC selection)	95.24 ± 5.56	76.84 ± 7.41	99.62 ± 1.23	92.44 ± 8.05	90.33 ± 11.52	66.67 ± 9.96	95.24 ± 8.57	67.32 ± 11.04
8	kPCA RBF + FDA (unsupervised PC selection)	87.86 ± 11.24	75.11 ± 15.02	99.12 ± 1.69	89.50 ± 9.41	87.45 ± 12.27	61.20 ± 12.91	94.78 ± 9.05	64.20 ± 11.19
9	kPCA RBF + FDA (supervised PC selection)	100.00 ± 0.00	64.07 ± 1.94	99.62 ± 0.92	92.02 ± 6.36	100.00 ± 0.00	69.49 ± 3.94	96.15 ± 7.29	58.13 ± 12.24
10[a]	P-RR	97.75 ± 2.36	83.88 ± 6.53						
11	P-PCR	91.25 ± 2.02	85.54 ± 4.45	99.47 ± 1.05	93.83 ± 3.23	92.69 ± 9.97	57.41 ± 8.80	93.33 ± 8.16	70.35 ± 8.19
12	P-ICR	93.63 ± 2.05	85.95 ± 5.16	98.08 ± 2.06	94.65 ± 2.89	83.32 ± 7.75	62.62 ± 5.84	91.10 ± 9.87	74.30 ± 7.30
13	PAM	91.50 ± 4.29	83.63 ± 5.82	99.74 ± 0.79	95.00 ± 4.61	89.35 ± 3.64	59.26 ± 9.22	98.57 ± 2.17	67.24 ± 6.58

[a]Because the gene numbers of the last three datasets are so great that our computer (CPU: 2 GHz Pentium IV) cannot process them using ridge regression model, we have not classified them using ridge regression model. On the contrary, ICA and PCA have the ability of compressing the gene expression data (as shown in Method Section), so we can use principal component regression or independent component regression to deal with large-scale data.

From Table 2 depicted above we can see that for colon, leukemia and glioma datasets, our proposed method is indeed efficient and feasible. Yet for hepatocellular data, the classification result of our method is not perfect. In addition, the other two methods we have used in our experiment (Methods 11 and 13) are even badly for this dataset. In fact, we can also find from Table 2 that, there is no method whose classification effect is always the best for all the four datasets.

The relationship between the credibility and the discrimination of eigenassay

Table 3 shows the credibility and its discrimination of every 30 eigenassay (strictly speaking, it is the corresponding eigengene's discrimination) extracted in one experiment (running ICA 100 times, deleting the eigenassays as described in the Experimental results section). We experimentalize using the colon data that the discrimination of every eigenassay is the accuracy on training set using leave-one-out cross-validation performance. Table 4 shows the 10 eigenassays corresponding to their credibility, which are orderly selected by SFFS algorithm during five experiments (using five random splittings of the colon data, running 100 ICA times, choosing 10 eigenassays using SFFS). Only from Table 3, we cannot find the certain relationship between the credibility and the discrimination of every eigenassay. Yet, from Table 4, we can see that most of the selected eigenassays have higher credibility. However, we also found from experiments that the higher credibility eigenassays are not all selected for classification. In addition, these results are also found in other experiments using other three datasets.

CONCLUSIONS

In this paper, we presented ICA methods for the classification of tumors based on microarray gene expression data. The methodology involves regularizing gene expression data using ICA, followed by the classification applying penalized discriminant method. We have compared the experimental results of our method with other 12 methods, which show that our method is effective and efficient in predicting normal and tumor samples from four human tissues. Furthermore, these results hold under re-randomization of the samples.

Because currently we have no suitable gene expression data of multiclass at hand, we only studied binary tumor classification problem in the experiments. In fact, our method is in essence a method that can address the problems with multi-classes. So we can use this novel method to solve those multi-classification problems directly.

We also found in experiment that compared with PCA, ICA is not always reproducible when used to analyze gene expression data. This problem had also been found by Chiappetta (2004) and Liebermeister (2002). In literature (Chiappetta *et al*., 2004), the authors had considered that the reason of this phenomenon is that the ICA algorithm may converge to local optima. In addition, they have given out a 'consensus source' search algorithm which yields extremely stable and robust estimates for the sources, as well as indications relative to their stability. In this paper, we also use this method to solve the unstability of ICA.

In future works, we will at large study the ICA model of gene expression data, how to apply the method proposed in this paper to

Table 3. The credibility and its discrimination of all 30 eigenassays

Eigenassay	Credibility(%)	Accuracy(%)
1	100	55.00
2	100	65.00
3	100	47.50
4	100	55.00
5	99	65.00
6	95	80.00
7	93	67.50
8	92	47.50
9	92	45.00
10	87	52.50
11	84	50.00
12	81	67.50
13	80	52.50
14	79	67.50
15	71	57.50
16	65	62.50
17	65	55.00
18	58	60.00
19	49	65.00
20	45	70.00
21	44	57.50
22	41	47.50
23	37	62.50
24	37	57.50
25	35	45.00
26	32	37.50
27	28	60.00
28	26	52.50
29	26	47.50
30	24	60.00

Table 4. The ten selected eigenassays and their credibility

No.1	Eigenassay	9	8	1	5	15	7	20	2	4	6
	Credibility(%)	95	95	100	98	69	98	52	100	99	98
No.2	Eigenassay	12	14	1	2	3	4	5	7	8	6
	Credibility(%)	84	83	100	100	100	100	99	96	93	97
No.3	Eigenassay	6	1	2	3	4	5	7	8	9	13
	Credibility(%)	98	100	100	100	100	99	97	95	94	86
No.4	Eigenassay	7	1	9	6	3	2	10	4	5	12
	Credibility(%)	93	100	91	95	100	100	85	100	98	79
No.5	Eigenassay	6	12	1	19	3	4	5	29	2	7
	Credibility(%)	95	81	100	49	100	100	99	26	100	93

solving multiclass problems of tumor classification and also study how to make full use of the information contained in the gene data to restrict ICA models so that more exact prediction of tumor class can be achieved.

ACKNOWLEDGEMENTS

The authors are grateful to Hong-Qiang Wang and Zhan-Li Sun for helpful discussions on this paper. This work was supported by the National Science Foundation of China (nos 30570368 and 60472111).

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bartlett,M.S. *et al*. (2002) Face recognition by independent component analysis. *IEEE Trans. Neural Netw*, **13**, 1450–1464.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Chiappetta,P. *et al*. (2004) Blind source separation and the analysis of microarray data. *J. Comput. Biol*., **11**, 1090–1109.

Comon,P. (1994) Independent component analysis—a new concept? *Signal Processing*, **36**, 287–314.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumor using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Feri,F.J. *et al*. (1994) Comparative study of techniques for large-scale feature selection. *In: Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems,* eds. ES Gelsema and LS Kanal. Amsterdam: Elsevier, 403-413.

Frank,I.E. and Friedman,J.H. (1993) A statistical view of some chemometric regression tools. *Technometrics*, **35**, 109–143.

Furey,T.S. *et al*. (2000) Support vector machines classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Ghosh,D. (2003) Penalized discriminant methods for the classification of tumors from microarray experiments. *Biometrics*, **59**, 992–1000.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hastie,T. *et al*. (1995) Penalized discriminant analysis by optimal scoring. *Ann. Stat*., **23**, 73–102.

Hastie,T. *et al*. (1994) Flexible discriminant analysis by optimal scoring. *J. Am. Stat. Assoc*., **89**, 1255–1270.

Hori,G., Inoue,M., Nishimura,S. and Nakahara,H. (2001) Blind gene classification based on ICA of microarray data. *Proc. 3rd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, SanDiego, USA. 332–336.

Hyvärinen,A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw*, **10**, 626–634.

Hyvärinen,A., Karhunen,J. and Oja,E. (2001) *Independent Component Analysis*. Wiley, NY.

Iizuka,N. *et al*. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, **361**, 923–929.

Lee,T.-W. *et al*. (1999) Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Comput*., **11**, 417–441.

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.

Martoglio,A-M. *et al*. (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, **18**, 1617–1624.

Massy,W.F. (1965) Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc*., **60**, 234–246.

Nutt,C.L. *et al*. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*., **63**, 1602–1607.

Pochet,N. *et al*. (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, **20**, 3185–3195.

Selaru,F.M. *et al*. (2002) Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology*, **122**, 606–613.

Shipp,M.A. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med*., **8**, 68–74.

Teschendorff,A.E *et al*. (2005) A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, **21**, 3025–3033.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

West,M. (2003) Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm. *Bayesian Stat*., **7**, 723–732.

West,M. *et al*. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Zheng,C.H. *et al*. (2005) Post-nonlinear blind source separation using neural networks with sandwiched structure. *LNCS*, **3497**, 478–483.

Zheng,C.H. *et al*. (2006) Nonnegative independent component analysis based on minimizing mutual information technique. *Neurocomputing*, **69**, 878–883.