Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 18 (2008) 741–750
Progress in Natural Science
www.elsevier.com/locate/pnsc

Gene association study with SVM, MLP and cross-validation for the diagnosis of diseases

Junying Zhang [a,*], Shenling Liu [a], Yue Wang [b]

[a] School of Computer Science and Engineering, Xidian University, 161 Postbox, Xi'an 710071, China
[b] Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Alexandria, VA 22314, USA

Received 14 November 2007; received in revised form 26 November 2007; accepted 26 November 2007

Abstract

Gene association study is one of the major challenges of biochip technology both for gene diagnosis where only a gene subset is responsible for some diseases, and for the treatment of the curse of dimensionality which occurs especially in DNA microarray datasets where there are more than thousands of genes and only a few number of experiments (samples). This paper presents a gene selection method by training linear support vector machine (SVM)/nonlinear MLP (multilayer perceptron) classifiers and testing them with cross-validation for finding a gene subset which is optimal/suboptimal for the diagnosis of binary/multiple disease types. Genes are selected with linear SVM classifier for the diagnosis of each binary disease types pair and tested by leave-one-out cross-validation; then, genes in the gene subset initialized by the union of them are deleted one by one by removing the gene which brings the greatest decrease of the generalization power, for samples, on the gene subset after removal, where generalization is measured by training MLPs with leave-one-out and leave-four-out cross-validations. The proposed method was tested with experiments on real DNA microarray MIT data and NCI data. The result shows that it outperforms conventional SNR method in the separability of the data with expression levels on selected genes. For real DNA microarray MIT/NCI data, which is composed of 7129/2308 effective genes with only 72/64 labeled samples belonging to 2/4 disease classes, only 11/6 genes are selected to be diagnostic genes. The selected genes are tested by the classification of samples on these genes with SVM/MLP with leave-one-out/both leave-one-out and leave-four-out cross-validations. The result of no misclassification indicates that the selected genes can be really considered as diagnostic genes for the diagnosis of the corresponding diseases.

Keywords: DNA microarray data; Curse of dimensionality; Gene selection; Diagnostic genes; SVM; MLP; Cross-validation

1. Introduction

The advent of DNA microarray data has proven to be of great practical assistance in the understanding of the pathogenetic mechanism, disease diagnosis, drug exploration and gene therapy, all on gene expression level [1]. Even though biologists have not had a comprehensive understanding about which genes induce which diseases, the advent of biochip, which is characterized with the high throughput of DNA microarray profiles, e.g., hundreds of thousands of genes' expression levels obtained in only one experiment, and more intelligent data processing approaches in data analysis with computers, may help to search for the relation between genes and diseases. In fact, biologists have already made all the possible genes which may cause some specific diseases into a biochip. Finding the cause of some specific diseases is what we call gene association study. Hence, for gene association study, what follows should be data analysis with computer and intelligent algorithms to search for the specific genes which cause the diseases. This new challenging task to computer science

* Corresponding author. Tel.: +86 13992815979; fax: +86 29 88201531.
E-mail address: jyzhang@mail.xidian.edu.cn (J. Zhang).

1002-0071/$ - see front matter © 2007 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2007.11.022

is referred to as diagnostic gene selection, or simply, gene selection.

By now, the study on the regulation between genes is just in its beginning, which makes it difficult to utilize biology-based relation between genes for understanding the cause of a disease in gene expression level. Hence, similar to almost all the related studies, in this study, genes are considered to be independent, i.e., each gene forms a dimension. Due to the high throughput of a biochip, this means that the dimensionality of the gene space, or feature space, is very high, and will be even higher with the development of gene microarray technology and the increase of biochip integration level. On the contrary, the number of samples (each sample corresponds to an experiment for a patient) is seriously limited. This is because it is impossible for scientists to collect so large enough patients/samples (more than 10 times of space dimensionality) [2,3] due to the cost of collection and some other reasons. Hence, the so huge number of dimensions and so the small number of samples in the space definitely lead to very serious "curse of dimensionality" [1,4]. For instance, given only 24 samples in a 12,000 dimensional space, we are required to perform disease detection, clustering, gene profiling of some specific cancer, as well as to identify genes or gene subsets that are sufficiently informative to distinguish between the 24 different-class samples with some generalization power. All these problems above generate a brand-new challenge to data processing: gene selection, i.e., to select those genes which are substantially responsible to the diagnosis of diseases. As a matter of fact, gene selection is to select a small subset of genes from thousands of genes such that it can maximize the separability among the different types of disease samples and therefore, be used as diagnostic genes for the diagnosis of the diseases. Two major problems are required to be solved: (1) How many genes should be selected? (2) Which genes should be selected? Taking into account that these two are NP-complete problems [5,6], one generally turns to work on a suboptimal subset of genes.

As a specific feature selection problem, gene selection has the following characteristics compared with a general feature selection problem: (1) the number of genes is much more than the number of samples, i.e., only a few samples (say, several tens) are in a huge dimensional space (say, thousands). This is absolutely different from the general feature selection problem where the dimensionality of the space is only tens while the number of samples in the space is hundreds or thousands; (2) gene selection is to select only a subset of genes (say 10 genes) from a huge number of genes (say 1000 genes), while general feature selection is to select a subset of features from a few features (e.g., only 12 features were selected from 18 features [6,7]); (3) misdiagnosis of a disease will be very costly and not permitted, and therefore, not only separability of different disease classes, but also generalization performance of the classifier is more required to be maximized. All the above characteristics make gene selection a brand-new and challenging topic in the related research area.

Recently, much work has been done on DNA microarray data analysis, such as clustering on gene data with hierarchical incremental neural network [8], gene selection with support vector machine [9,10], and biomarker identification with probabilistic principle component analysis [11]. One of the representative studies in gene selection was proposed by Javed Khan et al. [12], which deals with the classification of gene expression data from cDNA microarrays that contains 88 samples belonging to four cancer types (each sample has expression levels of 6567 genes). In the study, genes with over-low expression levels were filtered firstly, principal component analysis (PCA) was used for further reduction of the dimensionality from 2308 retained genes to 10, 3-fold cross-validation was then performed to train 1250×3 models of multilayer perceptron (MLP) with 10 inputs and 4 outputs (no hidden layer was included). After integrating these MLPs, a model with 2308 inputs (each corresponds to a gene) and 4 outputs (each corresponds to a cancer type) was obtained, followed by a gene-ranking procedure according to the sensitivity of the model's outputs to different genes. Finally, 96 genes were selected as diagnostic genes of the 4 cancer types. Since gene selection is evaluated with classification performance, gene selection and classification are closely related in this study, and also in our study.

Gene selection on DNA microarray data is extracting more and more interest in the analysis of DNA data. SNR method [13,14] is a representative method simply and widely used for gene selection by biologists, which we also used and compared with our proposed method in the experiment section of this paper. However, some studies [14] performed data classification directly after dimensionality reduction realized by PCA without any gene selection procedure at all; some algorithms were based on gene sensitivity [15], while Ref. [16] dealt with feature selection problem only for binary logical function, not for DNA data; the feature selection problem introduced in Refs. [7] and [17] only discussed situations where space dimensionality was far smaller than that of DNA microarray data (12 features were selected from only 18 features therein). The branch and bound algorithm and its various versions [6,7] cannot handle gene selection problem due to its huge dimensionality of genes at hand, either.

In this paper, we proposed an effective gene selection method, for discriminating multiple disease types with classifiers of not only strong separability but also strong generalization power among samples belonging to separate disease types for DNA microarray data. Experiments and analysis on real DNA microarray MIT data (two disease types) and NCI data (four disease types) validate the effectiveness of the proposed method in that only 11 genes (for MIT data) or 6 genes (for NCI data) are responsible for the related diseases, which will greatly reduce the size of the biochip for disease diagnosis and with an accurate prediction of unseen microarray samples for their disease types.

2. Gene association study based on SVMs/MLPs

Suppose m experiments were conducted on m patients with K diseases, leading to m gene expression level of samples in an n-dimensional gene space, each belonging to some disease type in the K disease types. Thus, the data for gene selection is $\{(x_j, y_j),\ x_j \in R^n,\ y_j \in \{0, 1, \ldots, K-1\},\ j = 1, 2, \ldots, m\}$, where n is the number of dimensions (genes) of the gene space, m is the number of samples and K is the number of disease types. In general, we have $n >> m$. In this section, we limit our study to $K = 2$, i.e., there are only two disease types for simplicity and therefore only a binary classifier is required for the classification of the samples belonging to these disease types.

2.1. Linear separability of a small number of samples in a huge dimensional space

For there are only a few number of samples in a huge dimensional gene space, i.e., $m << n$, and the samples belong to only two disease types, it is not difficult to imagine that only a linear classifier rather than a complicated nonlinear classifier is enough for separating all the samples in the space correctly. In fact, this is very easy to be guaranteed from the following theorem.

Theorem 1. *m samples belonging to two classes in an n-dimensional space, where $m \leqslant n$, can always be separated by a linear classifier if the samples are linearly independent.*

Proof. Without the loss of generality, suppose that the first m_1 samples belong to the first class (labeled with $y = 0$), and the rest $m - m_1$ samples belong to the second class (labeled with $y = 1$). Hence, we need to prove that if there exist a separation hyperplane l: (w^T, b) which satisfies

$$\begin{cases} w^{\mathrm{T}} x_i + b \geqslant 0 & i = 1, 2, \ldots, m_1 \\ w^{\mathrm{T}} x_i + b < 0 & i = m_1 + 1, \ldots, m \end{cases} \tag{1}$$

or equivalently

$$AW = B \tag{2}$$

where x_i is the ith sample, $W = (w^{\mathrm{T}}, b)_{\mathrm{T}}$, $A = \begin{pmatrix} x_1^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ x_m^{\mathrm{T}} & 1 \end{pmatrix}$, $B = (b_1, b_2, \ldots, b_{m_1}, -b_{m_1+1}, -b_{m_1+2}, \ldots - b_m)^{\mathrm{T}}$, and $b_i \geqslant 0$, $i = 1, 2, \ldots, m$. Since $x_1, x_2, \ldots, x_m$ are linearly independent, the rank of A is $r(A) = m$, and hence, we have $r(A, B) = r(A)$. Therefore, there must exists a separating hyperplane l:(w^{T}, b) for separating all the samples in the space correctly.

It is obvious that the condition in Theorem 1, i.e., "m samples in n-dimensional space ($m \leqslant n$) are linearly independent", is easy to be satisfied with respect to general DNA microarray data due to the fact that the number of samples is much smaller than the number of genes. As a result, a linear binary classifier can always be found to fulfill correct classification task for the binary disease classification problem in gene space.

2.2. Gene association study based on SVMs for two disease types

For classifying two disease types, only a few samples in a huge dimensional gene space will readily lead to the fact that the samples can be linearly separated from Theorem 1. Also, from the above proof process, the solution space is of $n + 1 - m$ dimensions due to $r(A, B) = r(A) = m < n + 1$, i.e., the solution can be expressed by $W = (w^{\mathrm{T}}, b)^{\mathrm{T}} = \eta + k_1\xi_1 + \cdots + k_{n+1-m}\xi_{n+1-m}$, where ξ_1, $\xi_2, \ldots, \xi_{n+1-m}$ are the basis of the corresponding equation $AW = 0$, η is a specific solution of Eq. (2), and $k_1, k_2, \ldots, k_{n+1-m}$ are arbitrary real numbers. This indicates that there are infinite solutions as decision boundaries for separating the binary classification problem. Notice that the objective of gene selection is to select a subset of genes, which cannot only separate the samples of the two disease types, but also with the largest generalization power.

Support vector machine (SVM), which is a learning algorithm [18], learns a linear or nonlinear binary classifier minimizing structural risk for obtaining the largest generalization power of the classifier. This makes it to be one of the best choices for gene association study and the classification of samples belonging to two disease types.

Obviously, it is reasonable to describe the separability of patterns and generalization performance of classifiers with a margin yielded by classification hyperplane. To our knowledge, with SVM, one can always find in gene space G the maximum-margin hyperplane: l:(w^{T}, b), which brings the margin of $\frac{2}{\|w\|}$ with the direction from the first class to the second class being $\frac{w}{\|w\|}$. Hence, if we represent the margin as a vector, the margin vector of l in G space is $M_G = \frac{2}{\|w\|} \cdot \frac{w}{\|w\|}$. Projecting l to an r-dimensional gene subspace G' consisting of any r genes, we get the projection of l in G space to be l':$(w'\mathrm{T}, b)$ in G' subspace, and

$$w_i' = \begin{cases} w_i & \text{if gene } i \text{ is selected} \\ 0 & \text{if gene } i \text{ is not selected} \end{cases} \qquad i = 1, 2, \ldots n \tag{3}$$

Our strategy for gene association study for two disease types is to select r genes from all genes such that the margin vector is obtained from the linear SVM, when projected to the r-dimensional gene subset defined by these r genes, i.e., $M_{G'} = \frac{2\|w'\|}{\|w\|^2}$, is the maximum. Notice that finding this maximal margin is equal to finding the largest r absolute elements of w. Therefore, selecting r genes simply falls into two steps: (i) training a linear SVM with data to obtain separating hyperplane (w^{T}, b); (ii) finding the r dimensions in which the absolute values of the elements of w are the largest. Then, the resultant r dimensions are the genes selected with this SVM-based approach.

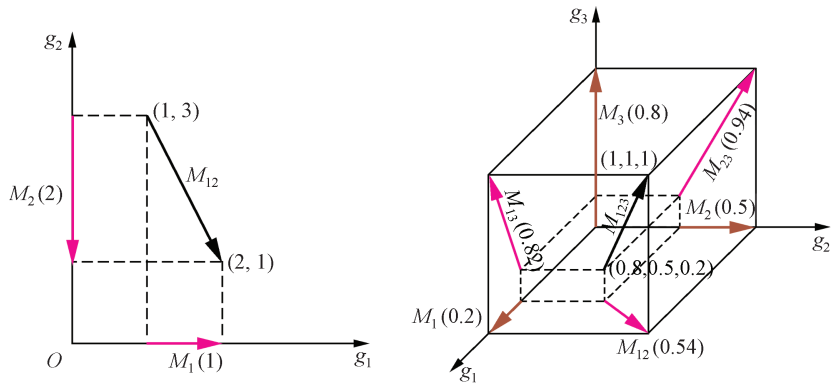

Fig. 1. Two demonstrations for SVM-based gene selection. (a) Selecting one gene from two; (b) selecting one or two genes from three.

Fig. 1(a) demonstrates a simple example in which altogether we have $n = 2$ genes and only one gene is selected, and Fig. 1(b) demonstrates the case we have $n = 3$ genes and one or two genes are selected. In Fig. 1, M_i indicates the margin vector when gene i is selected, M_{ij} indicates the one when genes i and j are selected, and M_{ijk} indicates the one when genes i,j,k are selected. It is certain from Fig. 1(a) that gene 2 should be selected due to the fact that it brings the maximal margin; similarly, from Fig. 1(b), gene 3 should be selected when only one gene is desired to be selected, and gene 2 and gene 3 should be selected when two genes are desired to be selected.

For any r-dimensional gene subspace, we always have $\|M_{G'}\| \leqslant \|M_G\|$. Since this approach is presented for the situation of a huge dimensional space and a small number of samples, it is not suitable for general feature selection.

2.3. Gene association study based on MLPs for multiple disease types

Now we consider the situation when there are multiple disease types in the dataset. Our aim is to select genes such that the samples belonging to separate disease types, if measured with the expression levels of the only selected genes, are the maximal discriminative and can be classified by a classifier with the largest generalization power. The main challenges of the problem include: (i) the number of disease types is usually more than two; (ii) the number of selected genes r is unknown in prior; and (iii) it is required that the generalization ability be the largest and the complexity be the lowest for the classifier defined on the selected genes.

Denote the sample set be C_i in which samples belong to the ith disease type. The idea of selecting genes in multiple disease types situation is as follows: search gene subset G_{ij} which can linearly separate samples in C_i and those in C_j, $i \neq j$ with the largest generalization power; get the union of all subset pairs G_{ij}, i.e., $G_0 = \cup_{i \neq j} G_{ij}$ (it is evident that the finally selected genes should be the subset of G_0); remove genes from the gene subset G_0, the one each time which mostly decreases the generalization power until any of genes could not be removed further, otherwise, the samples are not separable enough to be classified with some classifier.

Leave-K-out cross-validation is a general approach for understanding the generalization power of a classifier. Consider the situation that the dataset includes N samples. Leave-K-out cross-validation selects $N - K$ samples at random from the dataset as training data for the training of a classifier and uses the retained K samples as test data to test the generalization power of the classifier trained with the $N - K$ samples. Notice that K out of N classifiers in total, i.e., C_N^K classifiers, are required to be trained and tested for the approach, which may induce huge computations when N and K are large.

Support vector machine (SVM) and multilayer perceptron (MLP) are two typical classifiers which have been widely applied to many real world applications. The former is used as a linear or nonlinear binary classifier, and the latter as a nonlinear classifier in the study of microarray data here. We use SVM + leave-one-out and MLP + leave-one-out + leave-four-out to get the generalization power of SVM and MLP classifier, respectively, for selecting genes which are responsible for the separation of two disease types and multiple disease types, respectively. In our study, the ith SVM/MLP is denoted as $\text{SVM}_i/\text{MLP}_i$. Let misclassification rate of the ith classifier $\text{SVM}_i/\text{MLP}_i$ be $\delta_i(r)$, that is,

$$\delta_i(r) = \begin{cases} 100\% & \text{if there exist misclassified} \\ & \quad \text{samples in training dataset} \\ \delta_{ti}(r) & \text{otherwise} \end{cases}$$

where $\delta_{ti}(r)$ is the misclassification rate of test samples when r genes are selected, and $\delta_i(r)$ measures the generalization power both for r selected genes. Since altogether C_N^K SVM/MLP classifiers are required to be trained in leave-K-out cross-validation, the least generalization power of all these classifiers, i.e.,

$$\delta(r) = \max_{i=1,2,\ldots C_N^K} \delta_i(r) \tag{4}$$

can give us a comprehensive knowledge on the generalization power of the classifier with SVM/MLP structure, the

generalization power of it from arbitrary $N - K$ samples to the retained K samples.

According to the above discussion, we have the following procedure for selecting genes in multiple disease type situation.

1. Select genes such that samples belonging to any two different disease types are separable with the largest margin. This is performed via the steps that follow:

Step 1: Train a linear SVM with only the samples belonging to class i and class j to obtain the weight vector of the linear SVM classifier, w, with its ith element denoted as $w(i)$.

Step 2: Ranking genes in a decreasing order according to their values of $|w(i)|$, and set the number of selected genes r to be 1, i.e., $r = 1$.

Step 3: Select the first r genes as the selected gene subset, and train $C_N^1 = N$ linear SVMs, i.e., SVM_i, $i = 1, 2, \ldots, N$, with leave-one-out cross-validation for getting the generalization ability measure $\delta(r)$ calculated according to Eq. (4).

Step 4: If $\delta(r) \neq 0$, then $r \leftarrow r + 1$ and goto step 3; otherwise, store the selected gene subset to be G_{ij}, i.e., $G \leftarrow G_{ij}$.

2. Compute the union of all the gene subsets which can separate all class pairs G_{ij} and set it to be the candidate subset of selected genes, i.e., $G_0 = \cup_{i \neq j} G_{ij}$, and let $G \leftarrow G_0$, $r = |G|$.
3. Get the subset of genes which can separate all the classes. This is divided into the following steps:

Step 1: Remove only one gene, e.g., the jth gene g_j $(j = 1, 2, \ldots, r)$ from the gene subset G by performing MLP + leave-one-out in the retained $r - 1$ dimensional gene subspace and calculate the generalization ability measure denoted as $\delta^j(r)$ with Eq. (4), $j = 1, 2, \ldots, r$.

Step 2: Search for the gene k such that $\delta^j(r) \to \min$, and remove gene k from G, i.e., $G \leftarrow G - g_k$; let $r = |G|$, $\delta(r) \leftarrow \delta_k(r)$.

Step 3: Go to step 1 if $\delta(r) \neq 0$; otherwise we get the final selected gene subset to be G.

In the above algorithm, gene subsets, each for the separation of samples belonging to two different disease types, are at first extracted, with both the techniques of linear SVM classifier and leave-one-out cross-validation, and then the genes are deleted from the union of these gene subsets one by one by removing the gene which brings the greatest decrease of the generalization power of the MLP classifier for the samples on the gene subset after removal. Notice that both the two significant phases are computationally acceptable. In fact, for performance evaluation of the selected genes, we used the following measures in our experiments in the next section: (i) separability measure: the separability of a binary classifier is measured with Fisher/weighted Fisher criterion used in linear discriminatory analysis; (ii) generalization ability measure: this is measured with the maximal misclassification rate, obtained by linear binary SVMs with cross-validation for the classification of two disease types and/or nonlinear MLPs with cross-validation for the classification of multiple disease types.

3. Experiments and results

The experimental data herein include original gene expression data from patients of two leukemia types provided by MIT in USA (we call it MIT data here) and from samples of four tumor types provided by the National Cancer Institute of America (we call it NCI data here) [8]. We provide the experiments and comparison results of the former two disease types data due to their simplicity, and then put our focus on the latter multiple disease types data.

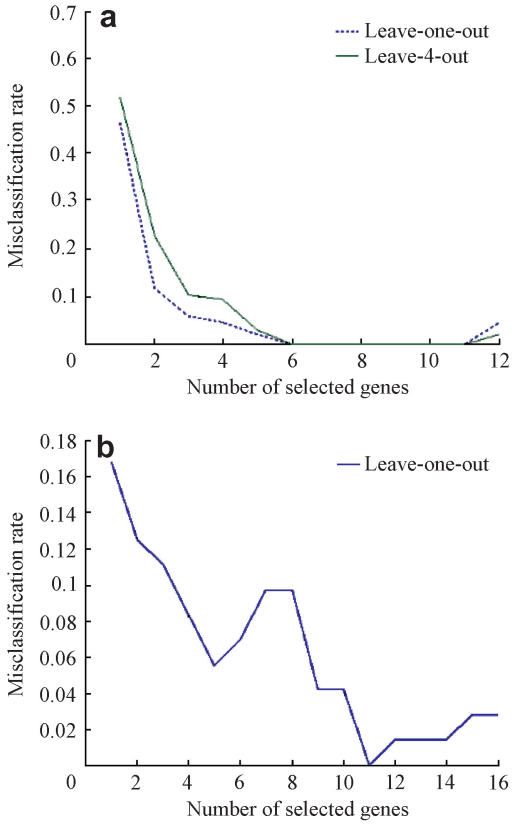

Fig. 2. (a) The maximal misclassification rate of 64 MLPs for leave-one-out cross-validation (dashed line) and that of $23 \times 8 \times 12 \times 21 = 46248$ MLPs for leave-four-out cross-validation vs. the number of selected genes, for NCI data; (b) the maximal misclassification rate of 72 SVMs for leave-one-out cross-validation vs. the number of selected genes, for MIT data.

MIT data are obtained from patients suffering from acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), composed of gene expression levels of 7129 genes with altogether 72 samples in which 47 are from ALL patient, and 25 are from AML patients. After gene selection procedure on MIT data, 72 SVM classifiers were trained with leave-one-out for the diagnosis of each binary disease types pair. The relationship of the maximal misclassification rate with the number of selected genes is illustrated in Fig. 2(b). As we can see, it is most appropriate to select 11 genes, with the corresponding maximal misclassification rate being 0. No curse of dimensionality occurs for this result, which will be discussed in detail later. We also performed a comparison with SNR gene selection method [12,13] in terms of generalization power, on the selected gene subset with the same dimensionality. Fisher and weighted Fisher (wFisher) values [19] were used as the criteria of generalization power too: the larger the Fisher/wFisher value is, the more separable the samples of different disease types will be. Table 1 illustrates the genes selected by the two methods (in the 3rd column) and the Fisher/wFisher values of samples on 11-dimensional selected gene subspace (in the 4th column). For visualizing the separability of samples on selected gene subspace, we projected gene expression data on selected gene subspace into the top two and top three principal components extracted by discriminate component analysis (DCA) [19] and weighted discriminate component analysis (wDCA) [19], and the Fisher/wFisher values of samples on 2-dimensional/3-dimensional projection space were calculated and are shown in the last 4 columns of Table 1. Fig. 3 shows the shatter plot of data after the projection from the selected gene subspace to the top 2/3 wDCA space, where Fisher refers to the value of Fisher Criterion, wFisher refers to the value of Weighted Fisher Criterion. Both Table 1 and Fig. 3 imply that the proposed method outperforms conventional SNR method in terms of separability of selected gene subset.

Table 2 shows the minimal margin with respect to the number of selected genes for the MIT data. It can be seen that the minimal margin for the case when 11 genes are selected is larger than those when 24, 30, 50, 100, 200 and 1000 genes are selected, respectively. Hence the generalization power of the samples reaches the maximum when 11 genes are selected, implying that these 11 genes should be considered as diagnostic genes for the diagnosis of ALL and AML.

NCI data come from 88 patients suffering from neuroblastoma (NB), rhabdomyosarcoma (MS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). Each microarray experiment was composed of gene expression levels of 2308 genes. Among the 88 samples, 64 samples were labeled, with 12/21/8/23 from NB/RMS/NHL/EWS patients, respectively, while all the other samples were not labeled. The experimental results and related analysis included the following: (1) gene selection and its result for the diagnosis of each pair of binary disease types; (2) the membership relation between gene subset for the diagnosis of each pair of binary disease types and that for the diagnosis of all related disease types; (3) gene selection and "curse of dimensionality"; (4) visualization of data on selected gene subspace for the knowledge of separability; (5) functions of hidden layer and output layer of an MLP classifier.

3.1. Gene selection for the diagnosis of two disease types

We perform gene selection for NCI data for each pair of disease types by the proposed method. Fig. 4 shows a part of gene selection procedure for separating each pair of disease types with SVM + leave-one-out. Ranking of genes in descending order according to the absolute value of weight coefficient of an SVM trained with samples belonging to disease types i and j [14,15] is done, and the best gene selection result G_{ij} in the sense of leave-one-out generalization is yielded when the first minimal non-negative margin appears with the rank of the genes. There is no necessity that more genes be selected, since the larger the minimal margin from SVM is, the better the classification performance is. As shown in Fig. 4(a), when $r = 5$, the minimal margin $= 0.065671 > 0$, and selected 5 genes (shown in the row of G_{12} in Table 3) allow SVM to classify all the

Table 1
Comparison of the proposed method and conventional SNR method for gene association study (MIT data)

Method	Genes (n)	Indices of selected genes	Fisher/wFisher of samples in 11-D gene subspace	Fisher/wFisher of samples in 11-D gene subspace projected to top 2 DCA space	Fisher/wFisher of samples in 11-D gene subspace projected to top 3 DCA space	Fisher/wFisher of samples in 11-D gene subspace projected to top 2 wDCA space	Fisher/wFisher of samples in 11-D gene subspace projected to top 3 wDCA space
SNR	11	1674, 1745, 1829, 1834, 2111, 2121, 2288, 3252, 3320, 4196, 4847	4.1142 17.7362	3.7681 16.4848	3.8119 16.6722	4.2786 17.2518	5.2057 18.9998
Proposed method	11	1779, 1868, 2402, 1882, 5952, 5710, 2345, 6218, 6179, 1763, 6201	6.6756 25.4741	8.9438 39.3764	9.6703 40.8678	8.5170 37.2610	8.5299 37.2739

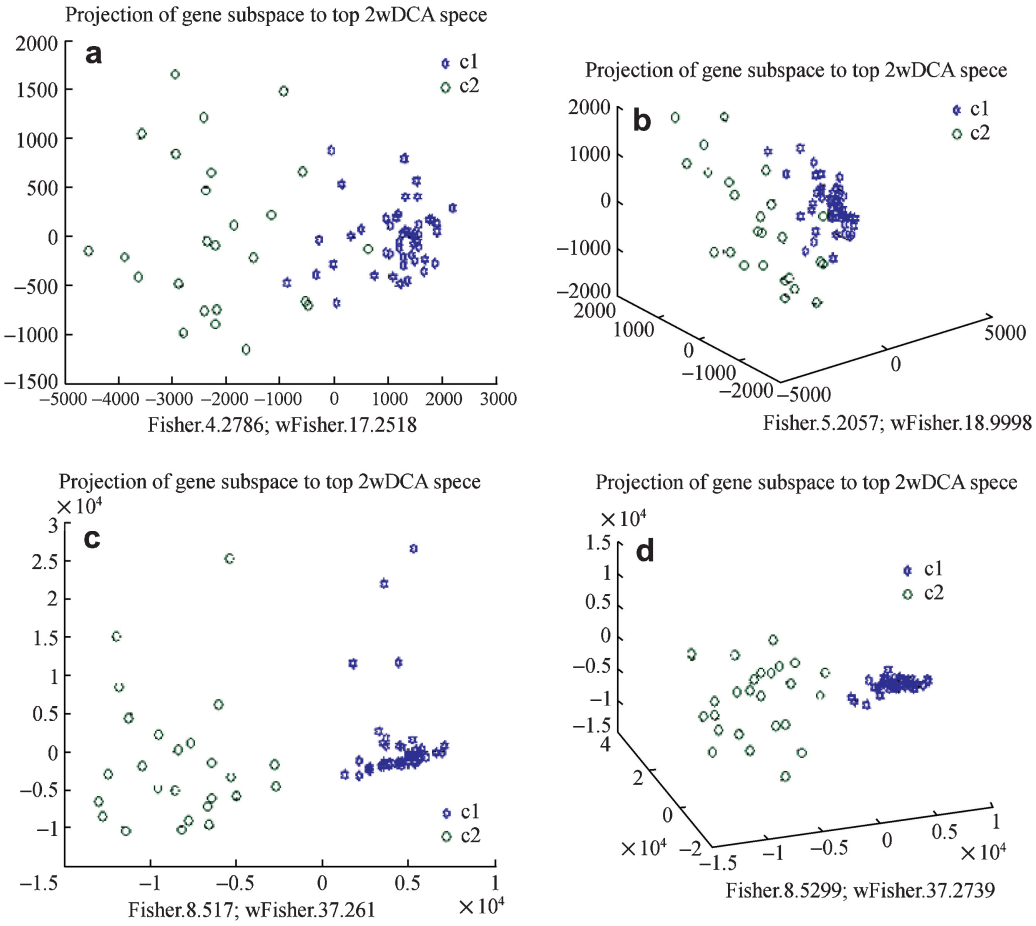

Fig. 3. Scatter plots of samples on the selected gene subset when projected to the top 2 (left column) and top 3 (right column) principal components with wDCA, with the conventional SNR method (upper figures) and with the proposed method (lower figures) for MIT dataset.

Table 2
The minimum margins for selecting different number of genes (MIT data)

Number of selected genes	11	24	30	50	100	200	1000
Corresponding minimum margin	1.999	1.945	1.692	1.669	1.483	1.358	1.309

samples correctly in the sense of leave-one-out cross-validation. But when r varies from 22 to 35, the minimal margin < 0, which will definitely lead to misclassification and the loss of generalization power. Similarly, the selected genes corresponding to Fig. 4(b)/(c) are shown in the row of G_{13}/G_{14} in Table 3, respectively. In Table 3, $G_{ij}(k)$ gives the selected gene subset for separating disease type i and disease type j, in which there are altogether k genes, with their gene indices shown in the row of $G_{ij}(k)$ in Table 3.

3.2. Membership relation of selected gene subsets

For further understanding the relation between the selected gene subsets for the separation of class pairs and of all classes, we sorted the selected genes according to the absolute value of the weight vector obtained from SVM for the separation of a class pair to measure the importance of the gene for this separation, which is expressed in the superscript of each selected gene shown in Table 3. We also divide the selected genes into 3-fold, 2-fold and 1-fold categories according to the overlap situation of the genes in the selected gene subset of G_{12}, G_{13}, G_{14}, G_{23}, G_{24}, G_{34}. For instance, in the row of $G_{12}(5)$ in Table 3, $246^3/1915^4/1750^2$ represent that the gene of the index 246/1915/1750 appears three times/twice/once in G_{12}, G_{13}, G_{14}, G_{23}, G_{24}, G_{34}, and rank at 3/4/2 in G_{12}, respectively (the upper the ranking position of a gene is, the more significant it is for classification). Then the question is if it is true that the genes which are of more overlap situation in separation of class pairs better for separation of all the classes. By removing genes from 20 genes in $G_0 = \cup_{i \neq j} G_{ij}$ one by one, we obtained only 6 genes with which the all 64 samples were classified without any misclassification for leave-one-out and leave-four-out cross-validations. This indicates that only 6 genes, i.e., the genes in the last row in Table 3, were left as diagnostic genes. Notice that the leave-one-out required 64 MLPs, while due to a huge computation in their training process for theoretical leave-four-out cross-validation which requires to train $C_{64}^4 = 15,249,024$ MLPs, we left one out from each class

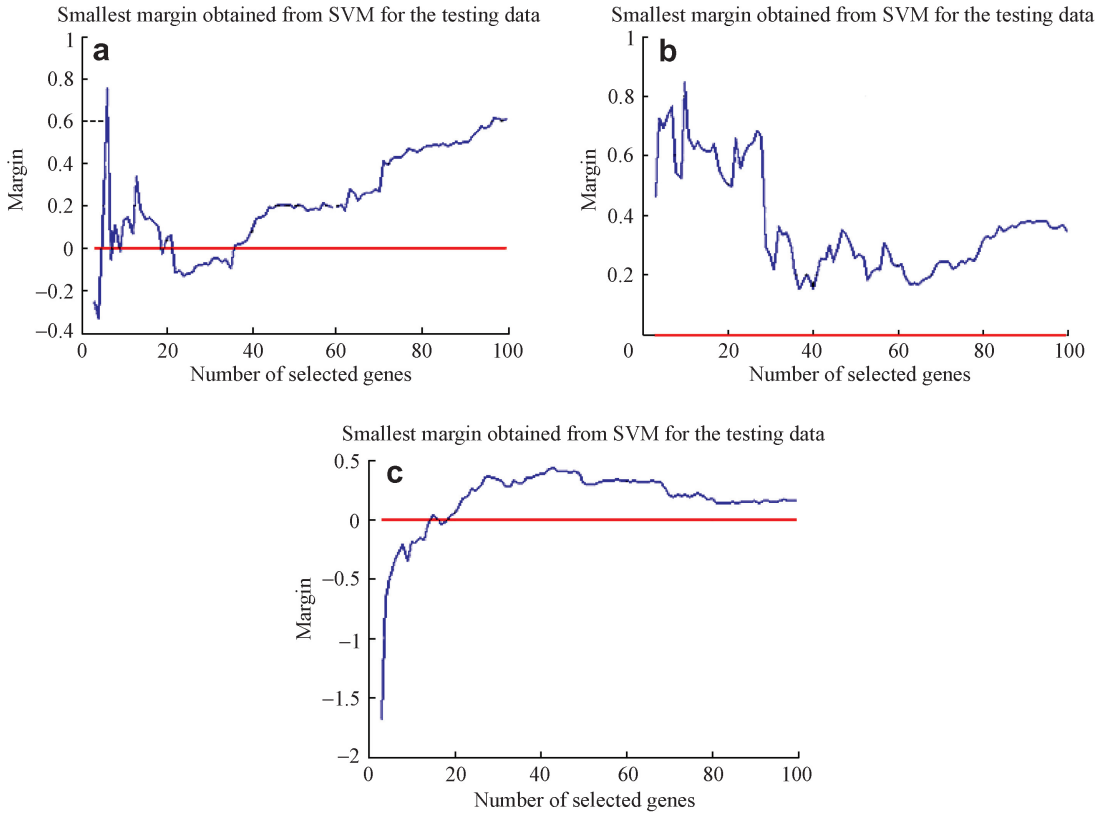

Fig. 4. The smallest margin of the linear SVM structure classifier among all the SVM classifiers, obtained from leave-one-out cross-validation, for searching genes for the separation of (a) disease type 1 and type 2; (b) disease type 1 and type 3; (c) disease type 1 and type 4, for NCI data.

Table 3
Selected gene subsets for the separation of disease type pairs and of all disease types (NCI dataset)

Gene subset	3-Fold genes				2-Fold genes			1-Fold genes	
G_{12} (5)	246^{3}	545^{5}	276^{1}					1915^{4}	1750^{2}
G_{13} (3)	246^{1}	545^{2}			1389^{3}				
G_{14} (15)	246^{3}	545^{7}		509^{1}	1389^{5}	1955^{4}	187^{2}		1954^{6}, 469^{8}, 1645^{9}, 2050^{10}, 1557^{11}, 1105^{12}, 135^{13}, 2046^{14}, 1093^{15}
G_{23} (3)			276^{2}					1915^{1}	151^{3}
G_{24} (3)			276^{3}	509^{1}			187^{2}		
G_{34} (3)				509^{2}		1955^{1}			523^{3}
G_{0} (20)	246	545	276	509	1389	1955	187	1915	1954, 469, 1645, 2050, 1557, 1105, 135, 2046, 1093, 1750, 151, 523

in the test data for our practical leave-four-out cross-validation, which requires to train only $23 \times 8 \times 12 \times 21 = 46248$ MLPs. The number of diagnostic genes obtained here is less than what we found before, 9 genes in Ref. [13] for this dataset.

From the above analysis, it is seen that the gene with more serious overlap may not be the final selected gene. As an example, genes 246, 545, 276 and 509 are all the 3-fold genes, while gene 276 was not finally selected but gene 151, which is only a 1-fold gene, was finally selected. This indicates that it is the union of these 6 selected genes that makes the best separation of all the 4 disease types.

3.3. Gene selection and the curse of dimensionality

Fig. 2 shows the relationship between the number of genes selected by removing a gene one by one from G_0 for G and the maximal misclassification rate of MLP + leave-one-out and leave-four-out/SVM + leave-one-out. As is seen, for NCI/MIT data, when 6–11/11 genes are selected, maximal misclassification rates of leave-one-out and leave-four-out/leave-one-out are 0, and these rates increase when the number of selected genes is too small or too large. This is just inconsistent with the "curse of dimensionality" mentioned in Refs. [1–4]. It is known from Ref. [4] that when the distribution of data is unknown in prior, the misclassification rate of a classifier with respect to n/N (n is the number of dimensions and N is the number of samples in the space) is shaped like a U curve. Generally speaking, "curve of dimensionality" does not occur when the number of samples N is more than 10 times the dimensionality of the space n. Evidently, only 64/72 samples in 2308/7129-dimensional gene space makes the curse of dimensionality very serious. However, the

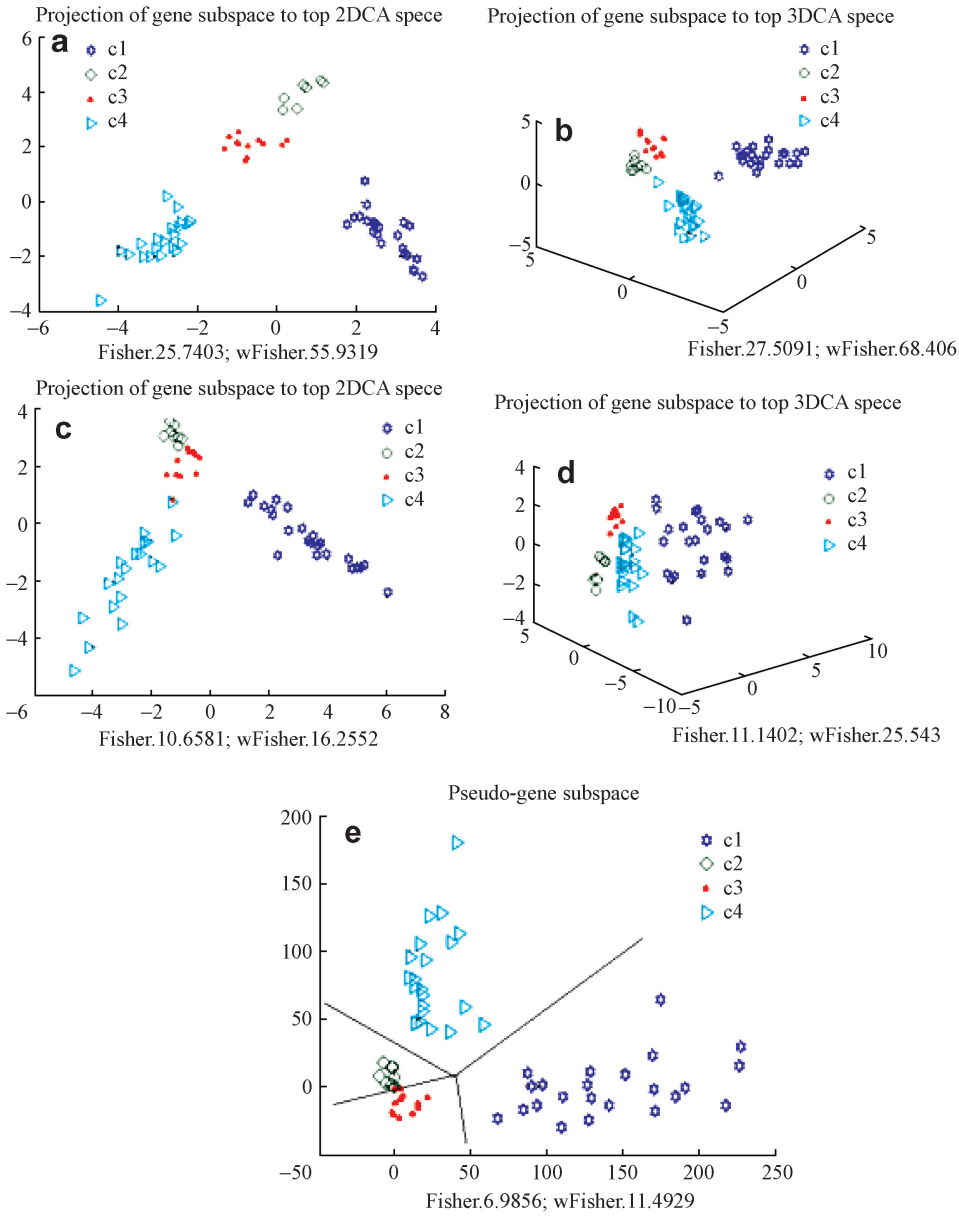

Fig. 5. Scattered plots of samples in (a) the initial gene subset G_0 and (b) the final gene subset G projected with DCA projection onto the top two (left column) and top three (right column) DCA principal components; (c) scattered plot of samples in MLP hidden space and the decision boundaries obtained from the output neurons of the MLP (for NCI dataset).

finally selected only 6/11 genes while 64/72 samples in this 6/11-dimensional gene subspace makes no curse of dimensionality.

3.4. Separability of data on selected gene subset

To visualize the separability of samples on initially selected gene subset G_0 and finally selected gene subset G for NCI data, we project the gene expression levels of all 4-class disease samples on G_0/G to top 2/3 principle components by discriminate component analysis (DCA) [19,20]. The left and right column of Fig. 5(a) and (b) show these projected results, respectively, demonstrating that the separability of samples does not deteriorate much when the 20 genes in G_0 is reduced to 6 genes in G: each class is still condensely clustered and classes are still well separated. This verifies that the selected 6 genes are in fact effective for the diagnosis of these 4 diseases, indicating the effectiveness of the proposed method.

3.5. Functions of hidden layer and output layer of an MLP

Since the number of selected genes is 6, and the number of disease types is 4, a 6-L-4 MLP, i.e., an MLP with 6 inputs, L hidden neurons and 4 outputs, was trained with the samples on the selected gene subset G. It has been recognized that the hidden layer performs a nonlinear map from input space to hidden space and the output layer realizes a linear classification of the projected data in the hidden space. Fig. 5(c) demonstrates the scattered plot of the

data on the hidden space of the trained 6-2-4 MLP after the nonlinear projection and the linear decision boundaries in the hidden space. By the comparison of the left figure in Fig. 5(b) with Fig. 5(c), it is recognized that the hidden neurons of the MLP realize a nonlinear projection which maps the samples belonging to all the disease types to sector regions originating from a common point in the hidden space: each disease type takes one of the sector regions and the sector region for a disease type is not overlapped with those of the other disease types. The output neurons generate decision boundaries which separate these sector regions by the learnt radioactive lines each of which passes this common point in the hidden space. Hence, sectorization (mapping the input data into sectors with different class within a different sector) and separating sectors with linear decision boundaries are the functions of the hidden neurons and output neurons, respectively, for an MLP to solve a nonlinear classification problem.

4. Conclusions

Gene association study is to discover the genes associated with the diagnosis and prediction of the susceptivity of diseases [21]. The advent of DNA biochip makes it possible to assay even thousands of genes simultaneously for understanding the substantial association of genes with complex diseases and/or cancers [11], while finding or discovering genes which are really responsible for diseases makes it great challenge to data processing scientists, including the curse of dimensionality which is especially serious in the DNA microarray data where the number of samples is considerably small compared with thousands of genes in a biochip – it is a vast combinatorial optimization problem, even to search for its suboptimal solution is a complex system engineering.

In this paper, a new method for gene association study is proposed based on SVM, MLP and cross-validation techniques. In the proposed method, linear binary SVMs and leave-one-out are used for selecting genes with best class-pair separation ability due to the best generalization performance of SVM, and MLPs, leave-one-out and leave-four-out cross-validation are used for finally obtaining diagnostic genes for the diagnosis of diseases. We applied our method to two real DNA microarray datasets, the datasets with the expression levels of 2308/7129 genes for 64/72 tissue samples belonging to 4/2 disease types. The 6/11 genes which preserved separability of the samples in different disease types well were finally selected for these datasets without any prior knowledge on how many genes are physically associated with the related diseases. Such 6/11 genes are diagnostic genes for the 4/2 diseases in that it is only required to make a biochip with 6/11 genes for the diagnosis of the 4/2 diseases for a tissue sample. This will make the cost of both biochip and diagnosis much lower than that for making a biochip with 2308/7129 genes, while still keeping correct diagnoses for the 4/2 diseases in an acceptable separability generalization sense.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 60574039, 60371044, 60071026), and the Sino-Italian joint cooperation foundation. The authors are indebted to the Computational Bioinformatics and Imaging Lab in Virginia Polytechnic Institute and State University in USA for providing us with the microarray data.

References

[1] Mjolsness E, DeCoste D. Machine learning for science: state of the art and future prospects. Science 2001;293(14):2051–5.
[2] Trunk GV. A problem of dimensionality: a simple example. IEEE Trans Pattern Anal Mach Intell 1995;1(3):306–7.
[3] Jain AK, Chandrasekaran B. Handbook of Statistics. Amsterdam: North-Holland; 1982.
[4] Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Prentice-Hall Inc.; 1999.
[5] Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.
[6] Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Trans Comput 1977;26(9):917–22.
[7] Hamamoto Y, Uchimura S, Matsunra Y, et al. Evaluation of the branch and bound algorithm for feature selection. Pattern Recognit Lett 1990;11:453–6.
[8] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001;17(2):126–36.
[9] Ramaswarmy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. PNAS 2001(26):15149–54.
[10] Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 2000;97(1):262–7.
[11] Xiong MM, Fang XZ, Zhao JY. Biomarker identification by feature wrappers. Genome Res 2001;11:188–1878.
[12] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576. Berkeley: University of California; 2000 [June].
[13] Zhang JY, Wang Y, Khan J, et al. Gene selection in class space for molecular classification of cancer. Sci China (Ser F) 2004;47(3):301–14.
[14] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673–9.
[15] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(3):389–422.
[16] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97(1–2):273–324.
[17] Jain A, Zongker F. Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 1997;19(2):153–8.
[18] Christopher JC. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov 1998;2:121–67.
[19] Loog M, Duin RPW. Multiclass linear dimension reduction by weighted pairwise fisher criteria. IEEE Trans Pattern Anal Mach Intell 2001;23(7):762–6.
[20] Anil K, Robert PR, Mar JC. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 2000;22(1):4–37.
[21] Hao K, Wang XB, Niu TH, et al. A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods. Human Molecular Genetics 2004;13(7):683–91.