BMC Bioinformatics

Methodology article

Gene selection for classification of microarray data based on the Bayes error

Ji-Gang Zhang[3] and Hong-Wen Deng*[1,2,3]

Address: [1]Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha, Hunan 410081, P. R. China, [2]The Key Laboratory of Biomedical Information Engineering of Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P. R. China and [3]Departments of Orthopedic Surgery and Basic Medical Science, School of Medicine, University of Missouri-Kansas City, 2411 Holmes Street, Kansas City, MO 64108, USA

Email: Ji-Gang Zhang - zhangjig@umkc.edu; Hong-Wen Deng* - dengh@umkc.edu

* Corresponding author

Published: 3 October 2007

BMC Bioinformatics 2007, **8**:370 doi:10.1186/1471-2105-8-370

Received: 21 March 2007
Accepted: 3 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/370

Abstract

Background: With DNA microarray data, selecting a compact subset of discriminative genes from thousands of genes is a critical step for accurate classification of phenotypes for, e.g., disease diagnosis. Several widely used gene selection methods often select top-ranked genes according to their individual discriminative power in classifying samples into distinct categories, without considering correlations among genes. A limitation of these gene selection methods is that they may result in gene sets with some redundancy and yield an unnecessary large number of candidate genes for classification analyses. Some latest studies show that incorporating gene to gene correlations into gene selection can remove redundant genes and improve classification accuracy.

Results: In this study, we propose a new method, Based Bayes error Filter (BBF), to select relevant genes and remove redundant genes in classification analyses of microarray data. The effectiveness and accuracy of this method is demonstrated through analyses of five publicly available microarray datasets. The results show that our gene selection method is capable of achieving better accuracies than previous studies, while being able to effectively select relevant genes, remove redundant genes and obtain efficient and small gene sets for sample classification purposes.

Conclusion: The proposed method can effectively identify a compact set of genes with high classification accuracy. This study also indicates that application of the Bayes error is a feasible and effective wayfor removing redundant genes in gene selection.

Background

One of the major applications of DNA microarray technology is to perform sample classification analyses between different disease phenotypes, for diagnostic and prognostic purposes [1-3]. The classification analyses involve a wide range of algorithms such as differential gene expression analyses, clustering analyses and supervised machine learning [4-6], etc. In classification analyses of microarray data, gene selection is one of the critical aspects [5,7-12]. Efficient gene selection can drastically ease computational burden of the subsequent classification task, and can yield a much smaller and more compact gene set without the loss of classification accuracy [13-17]. In addition, a smaller number of selected genes can be more conveniently and economically used for diagnostic purposes in clinical settings [18].

In the presence of thousands of genes in microarray experiments, it is common that a large number of genes are not informative for classification because they are either irrelevant or redundant [19]. Based on a review of the definitions of relevance [20,21], the genes can be classified into three disjoint categories, namely, strongly relevant, weakly relevant, and irrelevant genes [21]. Strong relevance indicates that the gene is always necessary for an optimal subset and cannot be removed without affecting the classification accuracy. Weak relevance indicates that the gene is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the gene is not necessary at all for classification accuracy. Hence, an optimal gene subset should include all strongly relevant genes, none of irrelevant genes, and a subset of weakly relevant genes. Based on the relevance definitions of genes, in classification analyses, a redundant gene is the one (which may be useful for classification analyses in isolation) does not provide much additional information if another informative gene is already present in the chosen gene subset.

Various approaches have been developed for gene selection to extract relevant genes from thousands of genes in microarray experiments, such as clustering methods [22] and pair-wise correlation analyses [2,6,23]. While multiple methods are available, it is well accepted in the field that a good gene selection method should be able to: 1) simplify the classifier by retaining only the relevant genes [23,24]; 2) improve or not significantly reduce the accuracy of the classifier; and 3) reduce the dimensionality of the dataset [9,25-27].

Traditionally, the methods for gene selection are broadly divided into three categories: filter, wrapper and embedded methods [21]. A filter method relies on general characteristics of the training data to select genes without involving any classifier for evaluation [10,28]. Many filter methods are usually mentioned as individual gene-ranking methods [1,10,26,29,30]. They evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. Although such gene ranking criteria are simple to use, they ignore correlation among genes, which may result in inclusion of redundant genes in selected gene set used for classification [31]. Redundant genes will increase the dimensionality of the selected gene set, and in turn affect the classification performance, especially on small samples [26]. In order to minimize redundant genes, correlation analyses have been incorporated in gene selection to remove redundant genes and improve classification accuracy [23,24,26,32]. The wrapper methods utilize the classifiers as evaluation functions and search for the optimal gene set for classification [14]. But the wrapper methods may suffer from excessive computational complexity. In contrast to the filter and wrapper approaches, the embedded methods perform the selection of genes during the training procedure and are specific to the particular learning algorithms [14,21,24,33]. For the wrapper and embedded methods, the search schemes are always involved to identify the optimal gene set for the sample classification. Searching the whole gene subset space may discover the optimal gene subset with respect to an evaluation criterion. However, an exhaustive search is usually computationally prohibitive. Thereby some partial search schemes are proposed, such as sequential forward selection, sequential floating forward selection, sequential backward elimination, sequential floating backward elimination and random search [34]. These partial search schemes are practically more feasible but provide no guarantee for identifying the optimal gene set [33].

In the earlier literatures, some excellent studies have highlighted the advantages of controlling classification error in yielding an optimal gene set, such as the study in the reference [35]. In this article we develop a novel gene selection method based on the Bayes error. Although the Bayes error has been used for feature selection in classification analyses, its use for gene selection in microarray data is very rare. It is well known that the Bayes error can provide the lowest achievable error rate bound for a given classification problem [36]. Theoretically, the Bayes error is the best criterion to evaluate effectiveness of gene set for classification [37], and the Bayes error depends only on the gene space, not the classifier itself [38]. From this point of view, by controlling the Bayes error it is feasible to find the optimal or sub-optimal gene set for a given classification problem without designing the classifiers. However, it is usually difficult to estimate directly the Bayes error rate analytically. An alternative is to estimate an upper bound of the Bayes error, which could be obtained by an error estimation equation based on the Bhattacharyya distance [37,39]. With this method, we can indirectly use the Bayes error for gene selection by controlling the upper bound of the Bayes error. This strategy is more promising than those requiring gene selection and classifier design simultaneously, as in the wrapper methods. Considering the promising aspects of the Bayes error, we propose in this study an approach, BBF (Based Bayes error Filter), for gene selection. Our selection algorithm is implemented in two steps: 1) first the relevant candidate genes are selected by a criterion function; and 2) the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes.

Application

To evaluate the performance of our proposed method in practice, we analyzed five publicly available microarray datasets: 1) Colon cancer dataset; 2) DLBCL dataset; 3)

Leukemia dataset; 4) Prostate dataset; 5) Lympho ma dataset (see Table 1).

Colon cancer dataset

This dataset consists of expression levels of 62 samples of which 40 samples are colon cancer samples and the remaining are normal samples [13,40]. Although originally expression levels for 6,000 genes are measured, 4,000 genes out of all the 6,000 genes were removed considering the reliability of measured values in the measured expression levels. The measured expression values of 2,000 genes are publicly available at [41].

DLBCL dataset

This dataset contains 77 samples in two classes, diffuse large B-cell lymphomas (DLBCL) and follicular lymphoma (FL), which have 58 and 19 samples, respectively [42]. The original dataset contains 7,129 genes. After the quality control, the dataset contains 77 samples and 6,285 genes. The measured expression values of genes are available at [30].

Leukemia dataset

This dataset, provided by Golub *et al.* [43], contains the expression levels of 7,129 genes for 27 patients of acute lymphoblastic leukemia (ALL) and 11 patients of acute myeloid leukemia (AML). After data preprocessing, 3,051 genes remain. The source of the 3,051 gene expression measurements is publicly available at [44].

Prostate dataset

This data set provides the expression levels of 12,600 genes for 50 normal tissues and 52 prostate cancer tissues [45]. The experiments were run on Affymetrix human 95Av2 arrays. The data preprocessing step leaves us with 6,033 genes. The data source is available at [46].

Lymphoma dataset

This dataset presented by Alizadeh *et al.* [47] comprises the expression levels of 4,026 genes. It contains 47 samples and two classes: germinal center B cell-like DLCL (diffuse large cell lymphoma) and active B cell-like DLCL. Among the 47 samples, 24 samples are germinal center B-like DLCL and 23 samples are active B cell-like DLCL. The dataset is available at [48].

Results

In the gene preselection step, we select the genes with FWER ≤ 0.05 (Family-Wise-Error rate). In our experiments, KNN and SVM classifiers are employed to demonstrate the proposed method and its classification performance. We choose the Euclidean distance in our KNN classifier with K = 5 and predict the class label by a majority vote. For the SVM classifier, we choose a linear kernel for decision plane computation.

We assess the performance of our method using the "Leave-One-Out Cross Validation" (LOOCV). LOOCV provides realistic assessment of classifiers which generalize well to new data [27]. The LOOCV method proceeds as follows: hold out one sample for testing while the remaining samples are used to make the gene selection and train the classifier. Note that to avoid selection bias [49], gene selection is performed using the training set. The genes are selected by our method using the training samples and then are used to classify the testing sample. The overall test error rate is calculated based on the incorrectness of the classification of each testing sample. Table 2 summarizes classification errors of five datasets with KNN and SVM classifiers by our method.

For the Colon dataset, as shown in Table 2, using the BBF method, 6 out of 62 samples are incorrectly classified by KNN and 8 by SVM, resulting in an overall error rate of 9.68% and 12.90%, respectively. According to the results of Ben-Dor *et al.* [13], without gene selection the classification error was 19.35% for KNN, and 22.58% for SVM, respectively. This is a significant improvement compared to the accuracy obtained by all available genes. This implies that there are irrelevant or redundant genes which deteriorate the performance of the classifiers, and the appropriate gene selection could effectively improve classification accuracy. The colon dataset has been used by many studies. For example, Liu *et al.* [23], used "normalized mutual information" with greedy selection and simulated annealing algorithm for gene selection. They reported that using KNN classifier the classification error is 9.68% with 29 selected genes for greedy selection and 12.90% with 26 selected genes for simulated annealing algorithm. Ding and Peng [26] proposed a "Minimum Redundancy – Maximum Relevancy" (MRMR) method.

Table 1: Summary for five datasets used in our experiments

Dataset	Source	No. of genes	No. of samples	Classes
Colon	Alon *et al.* (1999)	2000	62	Normal/Tumor
DLBCL	Shipp *et al.* (2002)	6285	77	DLBCL/FL
Leukemia	Golub *et al.* (1999)	3051	38	ALL/AML
Prostate	Singh *et al.* (2002)	6033	102	Normal/Tumor
Lymphoma	Alizadeh *et al.* (2000)	4026	47	Germinal/Activated

Table 2: The LOOCV errors for two-class datasets using KNN and SVM

Dataset	KNN(K = 5)		SVM(Linear)	
	Number of genes	Lowest Error (%)	Number of genes	Lowest Error (%)
Colon	12	9.68	20	12.90
DLBCL	6	7.79	5	9.09
Leukemia	3	0.00	2	0.00
Prostate	11	5.88	13	3.92
Lymphoma	8	2.13	3	0.00

Their best result by SVM was 8.06% with 20 genes, which means 5 out of 62 samples are incorrectly classified. Compared with our results, Liu's method [23] used more genes for similar classification errors. Using the SVM classifier, Ding and Peng [26] also selected 20 genes for best classification accuracy and the accuracy is slightly higher than ours. Some studies demonstrate that accurate diagnoses could be achieved using the expression levels of 15–20 genes from colon dataset [50]. To sum up the above results, the necessary number of genes could be less than 20 for the Colon dataset.

For the DLBCL dataset, in the original article, Shipp *et al*. [42] picked 30 genes by using their own weighted combination of informative genes. They correctly classified 71 out of 77 patients for a diagnostic error of 7.79%. Using our method 6 out of 77 samples are incorrectly classified by KNN and 7 by SVM, resulting in an overall error rate of 7.79% and 9.09%, respectively. However, only 5–6 genes are involved in classification and obtain similar classification error. Yang, *et al* [30] proposed GS1 and GS2 methods based on the ratio of inter-class and intra-class variation as a criterion function for gene selection. Using KNN they obtained classification error rate of 7.79% with 85 genes by GS1 and 6.49% with 70 genes by GS2; using SVM they achieved classification error rate of 3.90% for GS1 with 81 genes and GS2 with 55 genes. In contrast, we note that our results with KNN method are almost as good as theirs, yet only 6 genes are involved in the classification procedure and our classification error with SVM method is slightly higher than their results, but we use only 5 genes to reach similar level of performance.

For the Leukemia dataset, using the BBF method, all samples are correctly classified by KNN and SVM, with 3 and 2 genes for the two classifiers, respectively. Compared with other gene selection methods, our method appears to yield higher classification accuracy. For example, Dettling and Buhlmann [51] adopted four classifiers (LogitBoost, AdaBoost, KNN and Classification tree) to classify Leukemia dataset, and classification error was calculated based on LOOCV method. Their best classification error result was 1.39% with 25 genes by KNN. Some studies also come up with similar results to ours. Weston *et, al* [52] reported 0% classification error for a linear SVM using 20 genes by LOOCV method, but they used more genes.

For the Prostate dataset, in our results, 6 out of 102 samples are incorrectly classified by KNN and 4 by SVM, resulting in an overall error rate of 5.88% and 3.92%, respectively. Dettling and Buhlmann [22] proposed an algorithm for selecting supervised clusters of genes to find the gene groups for classification. They used KNN and aggregated trees methods, achieving the best results of classification error being 4.9% with 3 gene clusters by KNN. Our results are comparable to theirs, but used fewer genes. Gentile [53] used on a incremental large margin algorithm for gene selection and yielded 6.5% classification error estimated with 100 genes by LOOCV method. In contrast, we only used 11–13 genes and achieved better accuracy.

For the Lymphoma dataset, using our method, the classification error rates of KNN and SVM are 2.13% (this means only one sample is incorrectly classified) and 0%, respectively. Wang *et, al.* [54] also analyzed this dataset using several different gene selection methods and classifiers by the LOOCV method. In their study, they combined the locally linear embedding method and SVM classifier and yielded the classification error of 8.5% with 50 selected genes. When combining Signal-to-Noise method with KNN classifier, the classification error was 23.4%. The best error rate result they reported was 4.35% with 20 genes which is obtained by adopting the Information Gain method and a Neuro-Fuzzy Ensemble model. Diaz-Uriarte and Alvarez de Andres [55] reported the similar results to ours with the random forest method for gene selection, though they used a different estimation method for classification error rate. But they used more genes than our method. The results of our BBF method are comparable or outperform the above other results.

Discussion

As an important statistical index for classification analyses [56,57], the Bayes error has rarely been used in classification analyses of microarray data. In this study, we propose a novel gene selection approach for microarray classifica-

tion analyses. We introduce the Bayes error into the gene selection procedure, which turned out to be beneficial for classification analyses. The experimental results show that our proposed method can 1) reduce the dimension of microarray data by selecting relevant genes and excluding the redundant genes; and 2) improve or be comparable to the classification accuracy compared with other earlier studies.

In classification analyses, the classification error is an important criterion for selection of an optimal gene set. Some gene selection methods have been proposed to achieve the minimum classification error. Among these methods, a typical one is the method proposed by Peng *et al.* [35] which incorporates the classification error estimation with the gene selection method to determine an optimal gene set. The method first selects the genes with the highest relevance to the target classes, and minimizes the redundancy among the selected genes, and then determines an optimal gene set which has the minimum classification error estimated by cross-validation methods. In addition, the researchers presented the theoretical analysis [35]and comprehensive experimental studies [26] to prove that the criterion of "Maximum Relevance and minimum Redundancy" can benefit selection of optimal features for classification. Similar to the method of Peng *et al.* [35], our method uses the Bhattacharyya distance to select the genes with the highest joint relevance to the target classes, while minimizing the redundancy among the selected genes. Meanwhile, we can indirectly control the classification error due to the relationship between the Bhattacharyya distance and the Bayes error, and thus we can effectively avoid the computation of cross-validation error.

From the Bayesian decision theory, it is known that 1) the probability of error of any classifier is lower bounded by the Bayes error, 2) the Bayes error only depends on the gene space, not the classifier itself, and 3) there is always at least one classifier that achieves this lower bound [36,57,58]. Hence, from a theoretical point of view, it is possible to find out an optimal gene set for a given classification problem, rendering the minimum classification error. When selecting a set of relevant genes G with a minimum Bayes error in all gene space, according to the theories of the Bayes error, it can be guaranteed that at least one classifier may achieve this classification error. However, it should be noted that this optimal relevant gene set is classifier-specific. As pointed out [21,23], there are no relevancy definitions independent of the classifiers. That means not all classifiers can achieve the minimum Bayes error with the gene set G. This is because different classifiers have different biases and a gene which may be favorable for one classifier but may not for another. This phenomenon is also observed in our results. For example, in the colon dataset, for SVM classifier the best classification error was 12.90% with 20 genes, while for KNN classifier we could achieve the best classification error with 12 genes. When more genes are involved, the error rate for KNN classifier will increase. Since no single subset is optimal for all classifiers, it would be sensible to adopt a strategy to incorporate a classifier into gene selection for classification like a wrapper method. As the two-stage algorithm proposed in the early study [35], the first stage is to select relevant genes and eliminate redundant genes; the second stage is to search a more compact gene set for a specific classifier. This algorithm may not only yield an optimal or sub-optimal gene set for a specific classifier and increase the classification accuracy, but also decrease computation complexity when compared to a wrapper method. Our method can be extended to adopt this algorithm.

In classification analyses, genes obtained from observations may not be all informative for target classes. It is necessary to pick out the relevant candidate genes even though some of them are redundant. For our gene selection system, a gene preselection step is used to select the relevant candidate genes based on their individual relevance to the target classes. But the gene preselection step alone cannot yield the optimal gene set for a classification problem because it cannot eliminate the redundant genes due to the correlations between genes [4,5,36]. Efforts have been made to minimize the redundancy by measuring pair-wise gene correlations within the selected gene set [26,32,35,59], performing clustering analyses [23,26] and Markov blanket filtering [60]. Our proposed BBF method identifies the redundant genes by using Bhattacharyya distance measure to minimize the Bayes error. It is clear from Equation 2 that if a gene is highly correlated with another gene, combination of these two genes may not contribute more to the Bhattacharyya distance measure between two classes than any one of them. For an extreme example, i.e., the correlation of two genes is 1, it is impossible to calculate inverse matrix of covariance matrix for Bhattacharyya distance measure between two classes, and thus the redundant gene can be eliminated.

In addition, upper bound of the Bayes error (denoted by ε^*_B) is a critical parameter in this gene selection scheme. In the BBF method, we select genes according to their contribution to Bhattacharyya distance, d_B. It has been proved that ε^*_B monotonically increases as d_B increases, but at a decelerating manner since the rate of increase of ε^*_B decreases with the increasing of d_B [36]. When d_B increases to a certain level, e.g., 4.0, increasing d_B may not efficiently improve classification accuracy. Under this condition, with more genes selected, their contributions to classification accuracy turn out to be increasingly negligible. In the case of high ε^*_B, it may lead to a loss of some relevant

genes; in the case of smaller ε^*_B, it may involve some genes of negligible effects for classification. With this in mind, we set a criterion, ε^*_B being 1.0E-4, for picking relevant genes.

Conclusion

In summary, the BBF method can effectively perform gene selection with reasonably low classification error rates and a small number of selected genes. Our method may not only obtain a small subset of informative genes for classification analyses, but also provide a balance between selected gene set size and classification accuracy. This is confirmed by testing our method on the 5 real datasets.

Methods

In microarray classification analyses, the main objective of gene selection is to search for the genes which keep the maximum amount of information about the class and minimize the classification error. According to the Bayes theorem, the Bayes error can provide the lowest achievable error rate bound for a given classification problem (see details in [61,62]). The problem of gene selection is equivalent to determining the subset of genes which can minimize the Bayes error.

Let us consider the situation where a given gene expression measurement vector $\boldsymbol{x}$ needs to be classified into one of L classes. $P(c_i)$ denotes the a priori class probability of class i, $1 \le i \le L$, and $p(x|c_i)$ denotes the class likelihood, i.e., the conditional probability density of $\boldsymbol{x}$ given that it belongs to class i. The probability of $\boldsymbol{x}$ belonging to a specific class i, i.e., the posteriori probability $p(c_i|x)$, is given by the Bayes theorem:

$$p(c_i \mid x) = \frac{p(x \mid c_i)P(c_i)}{p(x)}$$

where $p(x)$ is the probability density function of x and is given by:

$$p(x) = \sum_{i=1}^{L} p(x \mid c_i)P(c_i)$$

When assigning a vector $\boldsymbol{x}$ to the class with the highest posterior probability, the error associated with this classification is called the Bayes error, which can be expressed as:

$$E_{bayes} = 1 - \sum_{i=1}^{L} \int_{C_i} P(c_i)p(x \mid c_i)dx$$

where C_i is the region where class i has the highest posterior. The probability of classification error of any classifier is lower bounded by the Bayes error [63,64]. However, the computation of the Bayes error is quite complicated. This is due to the fact that the Bayes error is obtained by integrating high-dimensional density functions in complex regions. Therefore, attention has focused on approximations and bounds for the Bayes error. One of these bound estimations for the Bayes error is provided by the Bhattacharyya distance.

In this study, we will use the Bhattacharyya distance to control the Bayes error for gene selection. For simplicity, we consider a binary classification study with m sample subjects and n measured genes. Assume there are two classes. Let x_{ij} be the expression measurement of the jth gene for the ith sample, where $j = 1, 2,..., n$, $i = 1, 2,..., m$. Here we assume $\boldsymbol{x_1},..., \boldsymbol{x_m}$ are the m samples, where $\boldsymbol{x_i} = [x_{i1}, x_{i2},..., x_{in}]$. Let $\boldsymbol{Y} = [y_1,..., y_m]^T$ denote the class labels of m samples, where $y_i = k$ indicates the sample i belonging to class k ($k = 1, 2$ stands for two different kinds of phenotypes, e.g., disease and control).

$$X = \begin{bmatrix} Gene1 & Gene2 & \cdots & Gene\mathbf{n} \\ x_{11} & x_{12} & & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \text{ and } Y = \begin{bmatrix} Label \\ y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

In general, gene selection is to select relevant genes and remove redundant genes. Hence our method is divided into two steps. The first step, *Gene preselection*, is to select relevant candidate genes. The second step, *Redundancy filter*, is to apply the criterion of controlling the upper bound of the Bayes error to the remaining genes obtained from the first step for eliminating the redundant genes.

Gene preselection

An intrinsic problem with microarray data is that sample size m is much smaller than the dimensionality of the genes. Our gene preselection is based on its strength for phenotype discrimination of each individual gene j, with $j \in \{1, 2,..., n\}$. We use a univariate criterion function (e.g. Wilcoxon test or F-test) to evaluate discriminative power for each gene:

$$Score(j) = S(j) \quad j \in \{1,..., n\} \tag{1}$$

where $S(\cdot)$ is the criterion function. We then select those relevant candidate genes according to FWER.

Redundancy filter

We use Bhattacharyya distance to estimate the upper bound of the Bayes error, which will be used as a criterion to filter out redundant genes from remaining genes derived from *Gene preselection step*. Before discussing this step, let us introduce the Bhattacharyya distance and the relationship between the Bhattacharyya distance and the

Bayes error. The Bhattacharyya distance, d_B, can be a separability measure between two classes and also can give lower and upper bounds of the Bayes error [36]. The Bhattacharyya distance is given as:

$$d_B = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\sum_1 + \sum_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2}\ln\frac{\left|\left(\sum_1 + \sum_2\right)/2\right|}{\left|\sum_1\right|^{1/2}\left|\sum_2\right|^{1/2}} \quad (2)$$

M_k is the mean vector of class k (k = 1 or 2); $\sum_k$ is the covariance matrix of class k (k = 1 or 2). The first term of Equation (2) gives the class separability due to the difference between class means, and the second term gives the class separability due to the difference between class covariance matrices. The Bayes error of classification between the two classes is bounded by the following expression:

$$\varepsilon_B \le \sqrt{P_1 P_2}\exp(-d_B) \quad (3)$$

where P_k is prior probability of class k (k = 1 or 2). We can derive the upper bound of the Bayes error evaluated from the Inequality (3) with $P_1 = P_2 = 0.5$. That is,

$$\varepsilon_B^* = 0.5\exp(-d_B) \quad (4)$$

We will use ε^*_B to control the Bayes error in classification analyses in order to filters out redundant genes. After the gene preselection procedure, the remaining genes form the candidate gene subset, B, which is regarded as an informative gene set. Then we construct an empty set, A, for selected relevant genes. We select the gene ranked first in the list of B as the initial gene in A. The gene ranked first is much more informative than any other genes to discriminate the two classes, thus we set it as the indispensable gene in A. We then use Sequential Forward Selection algorithm to select the genes with great contribution to the Bhattacharyya distance between two classes [34]. When the estimated upper bound of Bayes error reach predefined criterion, the search procedure stops and return the selected gene set, A.

In summary, our algorithm for gene selection proceeds as follows:

Step 1: Use a criterion function (we adopt Wilcoxon test in this study) to evaluate the discriminative power for each gene and select candidate genes according to FWER level.

Step 2: 1) Initialize A as an empty set (A is the set of selected relevant genes)

2) Initialize B as the candidate genes set, pick one gene ranked first in the list of criterion function values from B and put it into A as an initial gene.

3) For i = 2: t (t is the number of genes to be selected)

• for j = 1:q (q is the number of genes in B)

---Take gene j from B, put it into A, and calculate $d_B(j)$ with all genes in A.

• end

• Select the gene from B with the maximal d_B value and calculate corresponding ε^*_B

• if ε^*_B is greater than or equal to pre-defined criterion (here this criterion is set as 1.0E-4, which will be discussed later in Discussion section)

----Put this gene into set A; remove this gene from B

• else

----Stop the cycle and return the gene set A

4) End

Implementation of BBF method is available per request from zhangjig@umkc.edu with source code in R.

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

JGZ developed the algorithm and performed the analyses for five real-data sets. HWD provided discussion, general supervision and revised the manuscript for this study. Both authors have approved the final version of the manuscript.

Acknowledgements

The investigators of this work were partially supported by grants from NIH (R01 AR050496, R21 AG027110, K01 AR02170-01, and R01 AG026564-01A2). The study also benefited from grant support from National Science Foundation of China, Huo Ying Dong Education Foundation, HuNan Province, Xi'an Jiaotong University, and the Ministry of Education China. We thank Dr. YaoZhong Liu and Dr. YongJun Liu for proofreading the manuscript and discussion of some technical issues in this study.

References

1. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Am Stat Assoc* 2002, **97(457):**77-87.
2. Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C, Meltzer P: **Classification and diagnostic prediction of cancers using gene expression**

profiling and artificial neural networks. *Nature Medicine* 2001, **7(6):**673-679.
3. Lee Y, Lee CK: **Classification of multiple cancer types by multicategory support vector machines using gene expression data.** *Bioinformatics* 2003, **19(9):**1132-1139.
4. Baldi P, Long AD: **A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-test and Statistical Inferences of Gene Changes.** *Bioinformatics* 2001, **17:**509-519.
5. Li Y, Campbell C, Tipping M: **Bayesian automatic relevance determination algorithms for classifying gene expression data.** *Bioinformatics* 2002, **18:**1332-1339.
6. Varma S, Simon R: **Iterative class discovery and feature selection using Minimal Spanning Trees.** *BMC Bioinformatics* 2004, **5:**126.
7. Diaz-Uriarte R: **Supervised methods with genomic data: a review and cautionary view.** *Data analysis and visualization in genomics and proteomics* 2005:193-214.
8. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER: **Optimal number of features as a function of sample size for various classification rules.** *Bioinformatics* 2005, **21:**1509-1515.
9. Jirapech-Umpai T, Aitken S: **Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes.** *BMC Bioinformatics* 2005, **6:**148.
10. Lee JW, Lee JB, Park M, Song SH: **An extensive evaluation of recent classification tools applied to microarray data.** *Computation Statistics and Data Analysis* 2005, **48:**869-885.
11. Mukherjee S, Roberts SJ: **A Theoretical Analysis of Gene Selection.** *Proceedings of IEEE Computer Society Bioinformatics Conference (CSB 2004)* 2004:131-141.
12. Yeung KY, Bumgarner RE, Raftery AE: **Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data.** *Bioinformatics* 2005, **21:**2394-2402.
13. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *Proceedings of the fourth annual international Conference on Computational molecular biology* 2000:54-64.
14. Blanco R, Larranaga P, Inza I, Sierra B: **Gene selection for cancer classification using wrapper approaches.** *International Journal of Pattern Recognition and Artificial Intelligence* 2004, **18(8):**1373-1390.
15. Chow M, Moler I, Ejand M: **Identifying marker genes in transcription profiling data using a mixture of feature relevance experts.** *Physiol Genomics* 2001, **5:**99-111.
16. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2004:171-178.
17. Statnikov A, Aliferis C, Tsamardinos I, Hardin D, Levy S: **A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.** *Bioinformatics* 2005, **21(5):**631-643.
18. Tang EK, Suganthan PN, Yao X: **Gene selection algorithms for microarray data based on least squares support vector machine.** *BMC Bioinformatics* 2006, **7:**95.
19. Marchet A, Mocellin S, Belluco C, Ambrosi A, Francesco DeMarchi F, Mammano E, Digito M, Leon A, D'Arrigo A, Lise M, Nitti D: **Gene Expression Profile of Primary Gastric Cancer: Towards the Prediction of Lymph Node Status.** *Annals of Surgical Oncology* 2007:1058-1064.
20. Blum AL, Langley P: **Selection of relevant features and examples in machine learning.** *Intelligence* 1997, **97:**245-271.
21. Kohavi G, John R: **Wrappers for Feature Subset Selection.** *Artificial Intelligence* 1997:273-324.
22. Dettling M, Buhlmann P: **Supervised clustering of genes.** *Genome Biol* 2002, **3(12):**RESEARCH0069.
23. Liu X, Krishnan A, Mondry A: **An entropy-based gene selection method for cancer classification using microarray data.** *BMC Bioinformatics* 2005, **6:**76.
24. Ooi CH, Chetty M, Teng SW: **Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data.** *BMC Bioinformatics* 2006, **7:**320.
25. Dash M, Liu H: **Consistency-based search in feature selection.** *Artificial Intelligence* 2003, **151:**155-176.
26. Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data.** *J Bioinform Comput Biol* 2005, **3(2):**185-205.
27. Yu L, Liu H: **Redundancy based feature selection for microarray data.** *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* 2004:737-742.
28. Lai C, Reinders MJ, van't Veer LJ, Wessels LF: **A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets.** *BMC Bioinformatics* 2006, **7:**235.
29. Li T, Zhang C, Ogihara M: **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.** *Bioinformatics* 2004, **20:**2429-2437.
30. Yang K, Cai Z, Li J, Lin G: **A stable gene selection in microarray data analysis.** *BMC Bioinformatics* 2006, **7:**228.
31. Xiong M, Fang X, Zhao J: **Biomarker Identification by Feature Wrappers.** *Genome Research* 2001, **11:**1878-1887.
32. Xing E, Jordan M, Karp R: **Feature selection for high-dimensional genomic microarray data.** *International Conference on Machine Learning* 2001:601-608.
33. Tssamardinos I, Aliferis CF: **Towards Principled Feature Selection: Relevancy, Filters and Wrappers.** *Ninth International Workshop on Artificial Intelligence and Statistics* 2003.
34. Webb AR: **Statistical Pattern Recognition.** 2nd edition. London: Wiley, Chichester; 2002.
35. Peng HC, Long FH, Ding C: **Feature Selection Based on Mutual Information:Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.** *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 2005, **27(8):**1226-1238.
36. Lee C, Choi E: **Bayes error evaluation of the Gaussian ML classifier.** *IEEE Transactions on Geoscience and Remote Sensing* 2000, **38(3):**1471-1475.
37. Xuan GR, Zhu XM, Chai PQ, Zhang ZP, Shi YQ, Fu DD: **Feature Selection based on the Bhattacharyya Distance.** *18th International Conference on Pattern Recognition* 2006, **4:**957-960.
38. Carneiro G, Vasconcelos N: **Minimum Bayes Error Features for Visual Recognition by Sequential Feature Selection and Extraction.** *Proceedings of the Second Canadian Conference on Computer and Robot Vision* 2005:253-260.
39. Goudail F, Refregier P, Delyon G: **Bhattacharyya distance as a contrast parameter for statistical processing of noisy optical images.** *J Opt Soc Am A Opt Image Sci Vis* 2004, **21(7):**1231-1240.
40. Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine A: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proceedings of the National Accademy of Siences of the United States of America* 1999, **96(12):**6745-6750.
41. **Colon cancer dataset** [http://microarray.princeton.edu/oncology/]
42. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8(1):**68-74.
43. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: dass discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
44. **Leukemia dataset** [http://ligarto.org/rdiaz/Papers/rfVS/]
45. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1:**203-209.
46. **Prostate dataset** [http://ligarto.org/rdiaz/Papers/rfVS/]
47. Alizadeh AA, Eisen MB, Davis RE, Ma C, Losses IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
48. **Lymphoma dataset** [http://www.genome.wi.mit.edu/MPR]
49. Ambroise C, McLachlan G: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proceedings of the National Accademy of Siences of the United States of America* 2002, **99(10):**6562-6566.

50. Bo TH, Jonassen I: **New feature subset selection procedures for classification of expression profiles.** *Genome biology* 2002, **3:**.
51. Dettling M, Buhlmann P: **Boosting for tumor classification with gene expression data.** *Bioinformatics* 2003, **19(9):**1061-1069.
52. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V: **Feature Selection for SVMs.** *Advances in Neural Information Processing Systems* 2000 [http://www.cs.ucl.ac.uk/staff/M.Pontil/reading/featsel.pdf].
53. Gentile C: **Fast Feature Selection from Microarray Expression Data via Multiplicative Large Margin Algorithms.** *Proceedings NIPS* 2003 [http://books.nips.cc/papers/files/nips16/NIPS2003_AA2016.pdf].
54. Wang ZY, Palade V, Xu Y: **Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis.** *Proc of the Second International Symposium on Evolving Fuzzy System (EFS'06), IEEE Computational Intelligence Society* 2006:241-246.
55. Diaz-Uriarte R, Alvarez de Andres S: **Gene selection and classification of microarray data using random forest.** *BMC Bioinformatics* 2006, **7:**3.
56. Duda RO, Hart PE, Stork DG: **Pattern Classification.** second edition. Wiley, New York, NY; 2001.
57. Singh S, Kumar V, Singh M: **MULTIRESOLUTION ESTIMATES OF CLASSIFICATION COMPLEXITY AND MULTIPLE SUBSPACE CLASSIFIERS FOR UNDERSTANDING AND SOLVING COMPLEX RECOGNITION TASKS.** *Proceedings of the 24th IASTED International Multi-Conference* 2006:250-255.
58. Tumer K, Ghosh J: **Bayes Error Rate Estimation Using Classifier Ensembles.** *International Journal of Smart Engineering System Design* 2003, **5:**95-109.
59. Wang M, Wu P, Xia S: **Improving Performance of Gene Selection by Unsupervised Learning.** *Proceedings of Networks and Signal Processing* 2003, **1:**45-48.
60. Aliferis CF, Tsamardinos I, Statnikov A: **HITON: a novel Markov blanket algorithm for optimal variable selection.** *AMIA 2003 Annual Symposium Proceedings* 2003:21-25.
61. Devroye L, Gyorfi L, Lugosi G: **A Probabilistic Theory of Pattern Recognition.** Springer-Verlag New York, Inc; 1996.
62. Fukunaga K: **Introduction to Statistical Pattern Recognition.** Second edition. Academic Press, New York; 1990.
63. Devijver PA, Kittler J: **Pattern Recognition, a Statistical Approach.** Prentice Hall, Englewood Cliffs, London; 1982.
64. Schalkhoff R: **Pattern Recognition, statistical, structural and neural approaches.** John Wiley and Sons, New York; 1992.