BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1559–1564
doi:10.1093/bioinformatics/bti216

Gene expression

LS Bound based gene selection for DNA microarray data

Xin Zhou[1,2] and K. Z. Mao[1,*]

[1]School of Electrical and Electronic Engineering and [2]Bioinformatics Research Centre, Nanyang Technological University, Nanyang avenue, Singapore 639798

Received on October 28, 2004; accepted on December 9, 2004
Advance Access publication December 14, 2004

ABSTRACT

Motivation: One problem with discriminant analysis of DNA microarray data is that each sample is represented by quite a large number of genes, and many of them are irrelevant, insignificant or redundant to the discriminant problem at hand. Methods for selecting important genes are, therefore, of much significance in microarray data analysis. In the present study, a new criterion, called LS Bound measure, is proposed to address the gene selection problem. The LS Bound measure is derived from leave-one-out procedure of LS-SVMs (least squares support vector machines), and as the upper bound for leave-one-out classification results it reflects to some extent the generalization performance of gene subsets.

Results: We applied this LS Bound measure for gene selection on two benchmark microarray datasets: colon cancer and leukemia. We also compared the LS Bound measure with other evaluation criteria, including the well-known Fisher's ratio and Mahalanobis class separability measure, and other published gene selection algorithms, including Weighting factor and SVM Recursive Feature Elimination. The strength of the LS Bound measure is that it provides gene subsets leading to more accurate classification results than the filter method while its computational complexity is at the level of the filter method.

Availability: A companion website can be accessed at http://www.ntu.edu.sg/home5/pg02776030/lsbound/. The website contains: (1) the source code of the gene selection algorithm; (2) the complete set of tables and figures regarding the experimental study; (3) proof of the inequality (9).

Contact: ekzmao@ntu.edu.sg

1 INTRODUCTION

In recent years, research focus in molecular biology and genetics has shifted from the study of individual genes to the exploration of the entire genome. For example, the recently developed gene expression microarray technique measures the expression levels of thousands of genes in a single experiment. This large amount of data is something of a gold mine, from which a number of things can be found.

Gene-expression microarray data have been explored in a variety of ways including sample discriminant analysis and clustering, gene clustering and gene selection and many others. Among these issues, gene selection for discriminant analysis is of particular interest to us. From the viewpoint of discriminant analysis, we have a few reasons to perform gene selection. First, among the large set of genes many might be irrelevant, insignificant or redundant to a specific discriminant problem. Studies have shown that a small subset of genes might be sufficient for a particular biological problem (Golub *et al*., 1999). Second, gene selection reduces data volume and makes microarray data easier to handle and analyze. Third, reducing the number of genes decreases the demand for a large number of training samples because the performance of a pattern classifier partly depends on the ratio between the number of samples to the number of features. The collection of quite a large number of samples in microarray technique is expensive, time-consuming and even impossible. From the viewpoint of biologists, the importance of gene selection lies in its contribution to understanding diseases and functions of particular genes, and designing microarray experiments for clinical diagnosis and prognosis purpose.

In the context of gene-expression data analysis, several gene selection approaches were published. Golub *et al*. (1999) and Furey *et al*. (2000) employed an individual gene ranking score, Weighting factor, to perform gene selection prior to classification. Li *et al*. (2001) proposed a 'GA/KNN' method for gene assessment and sample classification. The main idea of GA/KNN is to find a huge number of optimal or near-optimal subsets and to assess the importance of genes for classification by examining the frequency of gene memberships in those subsets. Guyon *et al*. (2002) introduced a top–down recursive feature elimination (RFE) algorithm, in which features are successively eliminated during training of a sequence of support vector machine (SVM) classifiers.

If put in the context of pattern classification, gene selection can be solved as a feature selection problem. In general, a feature selection algorithm mainly consists of two basic components: search procedure and evaluation criterion (Dash and Liu, 1997). The search procedure generates candidate feature subsets for evaluation. In the search procedure, candidate feature subsets can be generated either sequentially or randomly. The well-known sequential forward selection (SFS) starts from an empty set and iteratively adds features, while the sequential backward elimination (SBE) starts from the full feature set and iteratively deletes features. The evaluation criterion measures the goodness of the candidate feature subsets generated by the search procedure. At each step of the iterative procedure, the feature that leads to the greatest improvement after addition or the least degradation after deletion is selected.

Generally, feature selection can be performed in two ways: the filter and the wrapper methods (Devijver and Kittler, 1982). The main difference between the two methods lies in the evaluation criterion. The filter method employs intrinsic properties of data, such as Mahalanobis class separability measure (Devijver and Kittler, 1982), as the criterion for feature subset evaluation, while the wrapper method evaluates feature subsets based on the performance of the

*To whom correspondence should be addressed.

© The Author 2004. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

classifier, such as classification errors. In most pattern recognition applications, the wrapper method outperforms the filter method. Several authors have employed wrapper methods in gene selection of microarray data (Inza *et al*., 2002; Xiong *et al*., 2001) and received satisfactory performance. However, the accuracy of wrapper methods is coupled with intensive computations. In contrast, the filter method is computationally efficient, and this makes the filter method very suitable for gene selection in high dimensional gene-expression data.

In our work we suggested a new filter-like evaluation criterion, called LS Bound measure, for gene selection. The new criterion has the advantages of both filter and wrapper methods. First, the criterion is derived from the leave-one-out cross validation (LOOCV) procedure of least squares support vector machines (LS-SVM) and is closely related to an upper bound of LOOCV classification results. Therefore, the criterion provides genes leading to accurate classification. Second, the estimation of the upper bound implicitly involves the training of the classifier only once, without repeated use of cross validation. As a result, the computational complexity is significantly reduced compared with the classical wrapper method. Our experiments showed that the computational complexity of our method is at the level of the filter method.

The paper is organized as follows. The LS-SVM is first briefly reviewed, and a new gene selection criterion, LS Bound measure, based on the upper bound for LOOCV of LS-SVMs is then proposed. The performance of the criterion is finally tested with two benchmark microarray datasets, i.e. the colon cancer dataset (Alon *et al*., 1999) and the leukemia dataset (Golub *et al*., 1999).

2 METHODS

2.1 Least squares support vector machines

In the past few years SVMs (Vapnik, 1998) have been introduced for solving pattern recognition problems. When used for classification, SVMs separate one class from the other with a hyperplane that maximizes the distance between the hyperplane and the nearest sample of each class. The determination of the hyperplane involves solving a quadratic programming(QP) problem, which requires expensive computation. To alleviate this problem, Suykens and Vandewalle (1999) proposed a least squares version of SVMs for classification problems. Instead of considering inequality constraints in the classical SVM approach, the LS-SVM employs equality constraints.

Consider l training data pairs: $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, l$, where $\mathbf{x}_i$ is an n-dimensional vector representing the i-th sample, and y_i is the class label of $\mathbf{x}_i$, which is either $+1$ or -1. The linear decision boundary is described as:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0, \tag{1}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_n]^{\mathrm{T}}$, and b is a scalar. In LS-SVMs, the optimization problem is formulated as:

$$\min_{\mathbf{w},\mathbf{e}} \Phi(\mathbf{w}, \mathbf{e}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \frac{\gamma}{2}\mathbf{e}^{\mathrm{T}}\mathbf{e}, \tag{2}$$

subject to equality constraints

$$y_i[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b] = 1 - e_i, \quad i = 1, \ldots, l, \tag{3}$$

where e_i denotes regression error for sample $\mathbf{x}_i$, $\mathbf{e} = [e_1, e_2, \ldots, e_l]^{\mathrm{T}}$, and γ is a given positive value assigned to penalize errors. The role of γ, just as that of the C in classical SVMs, is to adjust the compromise between generalization and training accuracy. The solution to the optimization problem is given by the saddle point of the Lagrangian:

$$L(\mathbf{w}, b, \mathbf{e}, \boldsymbol{\alpha}) = \Phi(\mathbf{w}, \mathbf{e}) - \sum_{i=1}^{l} \alpha_i [y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + e_i] \tag{4}$$

with Lagrange multipliers α_i. The conditions for optimality

$$\frac{\partial L}{\partial \mathbf{w}} = 0, \quad \frac{\partial L}{\partial b} = 0, \quad \frac{\partial L}{\partial \alpha_i} = 0 \quad \text{and} \quad \frac{\partial L}{\partial e_i} = 0$$

give

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \quad \alpha_i = \gamma e_i$$
$$\text{and} \quad y_i[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b] = 1 - e_i. \tag{5}$$

Equation (5) can be written in matrix form as:

$$\begin{bmatrix} 0 & -\mathbf{Y}^{\mathrm{T}} \\ \mathbf{Y} & \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{\mathbf{1}} \end{bmatrix} \tag{6}$$

where $\mathbf{Y} = [y_1, \ldots, y_l]^{\mathrm{T}}$, $\vec{\mathbf{1}} = [1, \ldots, 1]^{\mathrm{T}}$, $\alpha = [\alpha_1, \ldots, \alpha_l]^{\mathrm{T}}$ and $\boldsymbol{\Omega} = \{y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\}$.

Once the classifier is trained, we assign the corresponding class label of the test pattern $\mathbf{x}$ by the sign of the function $f(\mathbf{x})$:

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b. \tag{7}$$

2.2 LS Bound measure for gene selection

Feature selection approaches can be broadly grouped into filter and wrapper methods based on the evaluation criteria (Devijver and Kittler, 1982). The filter method evaluates feature subset based on intrinsic properties of data, which are related to the performance of the classifier but are not the direct function of the performance. In contrast, the wrapper method evaluates the feature subset based on the performance of the classifier directly. For better generalization, often the LOOCV error is employed to guide the selection in the wrapper method.

It is proved that the leave-one-out procedure gives an almost unbiased estimate of the probability of test error (Luntz and Brailovsky, 1969). But to obtain the leave-one-out error of each training data for a particular gene set, it requires repeated training of classifiers in the leave-one-out procedure, which makes the procedure a burdensome task. Several bounds on the expectation of SVMs from the leave-one-out estimator were introduced to reduce the high computational complexity in the leave-one-out procedure (Vapnik and Chapelle, 2000). SVMs are well suited to work with high dimensional data, such as microarray data (Furey *et al*., 2000). But the determination of the hyperplane involves solving a QP problem, which requires expensive computation. To alleviate this problem, our present work is focus on the LS-SVMs (Suykens and Vandewalle, 1999). The advantage of LS-SVMs is that the solution of LS-SVM can be obtained from solving a set of linear equations, which is much easier than the QP in the classical SVM. Although the performance of LS-SVMs on the classification of gene-expression data might not be better than classical SVMs, the simplicity and high computational efficiency of LS-SVMs will be of great benefit to the selection of gene subset from the high dimensional gene-expression data.

The decision function of a linear classifier can be written as:

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b. \tag{8}$$

The classification is based on the sign of the decision function $f(\mathbf{x})$ in Equation (8). In the present work, we are concerned with value of $f(\mathbf{x})$ rather than the sign of $f(\mathbf{x})$, because the value of $f(\mathbf{x})$ not only indicates the information of the class label of the sample $\mathbf{x}$, but also indicates the distance of sample $\mathbf{x}$ to the decision boundary. This value will change whenever a feature is added to the feature subset or deleted from the feature subset. Even if different feature subsets lead to the same classification results, the value $f(\mathbf{x})$ also provides a discriminant power to identify which subset is the best.

Next, we present a leave-one-out bound of $f(\mathbf{x})$ by training once on the entire training set to avoid repeated training for the leave-one-out procedure. It can be proved (see proof in the companion website) that when LS-SVM classifier is trained on the entire training set, if the corresponding Lagrangian

multiplier (α_p^0) of the training sample $\mathbf{x}_p$ is positive, the following inequality holds in the leave-one-out procedure.

$$-y_p f^p(\mathbf{x}_p) \leq \alpha_p^0[(D_{\min}^p)^2 + 2/\gamma] - 1, \tag{9}$$

where f^p is the decision function given by the LS-SVM after the sample $\mathbf{x}_p$ has been removed, α_p^0 is the corresponding Lagrangian multiplier of $\mathbf{x}_p$ when the LS-SVM classifier is trained on the entire training set, and $D_{\min}^p$ is the distance between $\mathbf{x}_p$ and its nearest neighbor. $f^p(\mathbf{x}_p)$ is the validation result for the sample $\mathbf{x}_p$ in the leave-one-out procedure. If $y_p f^p(\mathbf{x}_p)$ is negative the sample $\mathbf{x}_p$ is considered as a leave-one-out error, and if $y_p f^p(\mathbf{x}_p)$ is positive $\mathbf{x}_p$ is correctly classified in the leave-one-out procedure.

In the LS-SVM, we have $\alpha_p^0 = \gamma e_p^0$ and $y_p f^0(\mathbf{x}_p) = 1 - e_p^0$ [see Equations (5) and (3)]. According to the two equations, we group the training data into three categories based on the value of Lagrange multiplier α^0. The first category includes the samples with $\alpha_i^0 \leq 0$, which are far away from the decision boundary. They could be correctly classified not only when the decision function is trained on the basis of the entire data set, but also during the leave-one-out cross validation procedure in normal case. The samples with $0 < \alpha_i^0 < \gamma$ are in the second category. Although these samples are correctly classified in the training stage, they might be misclassified during the leave-one-out procedure because they are close to the decision boundary. The third category comprises samples with $\alpha_i^0 \geq \gamma$, which are misclassified not only in the training procedure but also in the leave-one-out procedure. The bound on the right side of inequality (9) is especially useful for the data in the second category. The upper bound is related to both the corresponding training result and the nearest neighbor. If the bound is negative the sample must be correctly classified in the leave-one-out procedure. If the bound is positive, although we are uncertain whether it will be misclassified in the leave-one-out procedure, the bound indicates to some extent the probability of misclassification, and hence can be used to evaluate the goodness of the feature (gene) set.

Combining the bounds for all training data together, we propose the following measure, called LS Bound measure, as the evaluation criterion for gene selection:

$$\mathbf{M} = \sum_{p=1}^{l} (\alpha_p^0[(D_{\min}^p)^2 + 2/\gamma] - 1)_+, \tag{10}$$

where $(x)_+ = \max(0, x)$. α^0 can be obtained from solving a set of linear equations in the LS-SVM. The solution of α also can be simply formulated as:

$$\alpha = \mathbf{H}^{-1}\vec{\mathbf{1}} - (\mathbf{Y}^{\mathrm{T}}\mathbf{H}^{-1}\vec{\mathbf{1}})(\mathbf{H}^{-1}\mathbf{Y})(\mathbf{Y}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Y}) \tag{11}$$

where $\mathbf{Y} = [y_1, \ldots, y_l]^{\mathrm{T}}$, $\vec{\mathbf{1}} = [1, \ldots, 1]^{\mathrm{T}}$ and $\mathbf{H} = \mathbf{\Omega} + \gamma^{-1}\mathbf{I}$. Note that the matrix $\mathbf{H}$ is positive definite and therefore invertible. The measure M gives a bound on error expectation, and can be considered as an estimate of the generalization performance. The feature (gene) subset which minimizes the measure M is preferred.

The LS Bound measure M in Equation (10) can be combined with any search algorithm, such as the sequential forward selection (SFS), to form a gene selection algorithm. The SFS algorithm is a simple greedy heuristic search algorithm. For better performance, other complex search algorithms, such as sequential floating forward selection (SFFS) (Pudil *et al*., 1994), can be used but at the cost of increasing the computational complexity. In this paper we only focus on the SFS algorithm, while the implementation and experimental results of combining LS Bound with SFFS algorithm can be found in our website (http://www.ntu.edu.sg/home5/pg02776030/lsbound/).

The pseudo code of the sequential forward gene selection algorithm can be summarized as follows:

The gene selection algorithm combining the LS Bound measure M with SFS

```
(1) Initialize S to an empty set;
     /* S is the set of selected genes */
(2) Initialize C to the full gene set;
     /* C is the set of candidate genes
     for selection */
(3) For i = 1 to m
     /* m genes are expected to be
     selected */
     p = number of genes in set C;
     For j = 1 to p
          Take gene j from set C and
          temporarily put into set S;
         Calculate the measure M using
         all genes in set S;
      End
      Select the gene with the minimal M;
      Put the selected gene into set S;
    End
```

The stopping criterion in our gene selection algorithm is whether a predefined number of features are selected. In our experiments, we first selected 50 genes from the original gene set. However, the 50 genes are not the final selection result because few of them can get the same or satisfactory performance. In our opinion, the LS Bound measure reflects the generalization performance of selected gene subset. The decrease of the LS Bound measure indicates the improvement of the generalization performance. The decision of final gene subset can be made based on whether the LS Bound measure approaches the minimum, or whether adding more genes results in insignificant change. For example, on the colon cancer dataset (see Results section), we selected 15 genes because adding more genes would not reduce the LS Bound measure significantly. For well separated dataset, after selecting several genes the LS Bound measure would be reduced to zero. Under such a case, the gene selection may be terminated when LS Bound measure is very close to zero.

To select the relevant genes using the LS Bound measure proposed in the present study, it is important to find the appropriate value of γ in Equation (10). This can be done using a simple technique. We employed our algorithm with a sequence of given values of γ to select important genes on the entire dataset. Because the LS Bound measure indicates the generalization performance, the optimal value of γ is chosen to be the one which gives the minimal LS Bound measure during the selection procedure.

2.3 Efficient computation of LS Bound measure

In the gene selection algorithm described above, the LS Bound measure M needs to be computed for every candidate gene set. For a high dimensional dataset such as microarray data, the repeated computation of M involving the inverse of an l-by-l matrix $\mathbf{H}$ in Equation (11) is actually a computationally intensive task. However, in the SFS search procedure we can express the inverse by using the Sherman–Morrison–Woodbury formula (Golub and van Loan, 1996):

$$(M + BCD^{\mathrm{T}})^{-1} = M^{-1} - M^{-1}B(C^{-1} + D^{\mathrm{T}}M^{-1}B)^{-1} \times D^{\mathrm{T}}M^{-1}. \tag{12}$$

Suppose that we have selected a particular gene set S and the matrix H for gene set S is denoted by $\mathbf{H}_S$. During the SFS search procedure, we have to temporarily put each of the candidate genes into gene set S to test the performance of each candidate gene. When gene k is added into gene set S to form a temporary set $S' = S \cup \{x_k\}$, the matrix H for gene set S', denoted by $\mathbf{H}_{S'}$, can be formulated as

$$\mathbf{H}_{S'} = \mathbf{H}_S + \mathbf{z}\mathbf{z}^{\mathrm{T}}, \tag{13}$$

where vector $\mathbf{z} = (y_1 x_{1k}, \ldots, y_l x_{lk})^{\mathrm{T}}$. According to the Sherman–Morrision–Woodbury formula, we have:

$$\mathbf{H}_{S'}^{-1} = \mathbf{H}_S^{-1} - \frac{\mathbf{H}_S^{-1}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{H}_S^{-1}}{1 + \mathbf{z}^{\mathrm{T}}\mathbf{H}_S^{-1}\mathbf{z}}. \tag{14}$$

Equation (14) reveals that the inverse of $\mathbf{H}_{S'}$ corresponding to a new subset $S' = S \cup \{x_k\}$ can be recursively computed from the inverse of $\mathbf{H}_S$ using simple matrix operations. Thus, in the SFS procedure, the complex inverse operation

is no longer needed by using the trick of Equation (14). Even compared with the filter measures, our measure, as well as our algorithm, is quite efficient. For example, 33 s are needed to select 50 genes from 1000 candidate genes using our gene selection algorithm on the colon cancer dataset in the Matlab environment (2.5 GHz P4 CPU with 512 MB RAM), while 52 s are needed in the Mahalanobis class separability measure-based sequential forward gene selection.

3 RESULTS

In this section we report the performance of the proposed gene selection measure and algorithm on two publicly available microarray datasets: colon cancer (Alon *et al*., 1999), and leukemia (Golub *et al*., 1999). Each of these datasets was pre-processed using the procedure described in Dudoit *et al*. (2002). After thresholding, filtering and logarithmic-transforming, the microarray data were standardized to zero mean and unit standard deviation across genes. Because the dimensionality (number of genes) of microarray data is huge, and many of the genes are irrelevant to the discriminant task, we employed a pre-selection procedure to reduce the searching space and computational time. We selected top 1000 genes based on Fisher's ratio, $f = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$. All the simulations and comparisons in this paper are based on the pre-processed and pre-selected data.

There are two goals in this section. One is to identify (select) genes which are significantly more important than others for the classification. Because of the small sample size of microarray dataset the goal can be fulfilled by selecting gene subset on the basis of the entire sample data to make the selection more reliable.

The other goal is to evaluate the performance of our gene selection criterion and algorithm, and to compare the performance with other gene selection algorithms. To assess the performance of the gene selection algorithm, in the literature some authors randomly split the original dataset into two sets, a training set and a test set, and the gene selection procedure was performed based on the training set while the performance of selected genes was assessed from the test set. Due to the small sample size of gene-expression data, however, the approach is not advisable. While some employed the entire dataset for gene selection, the performance of selected genes was tested using k-fold cross validation. This kind of cross validation, called internal cross validation, produces biased estimate. Ambroise and McLachlan (2002) suggested techniques of external (10-fold) cross validation and external .632+ Bootstrap, in which the gene selection and validation are performed on the different parts of the sample set, to obtain an unbiased estimate. Considering the high variance problem of cross validation, especially for the small-sample microarray data (Braga-Neto and Dougherty, 2004), we employed the external .632+ Bootstrap (Ambroise and McLachlan, 2002; Efron and Tibshirani, 1997) to evaluate the performance of our gene selection algorithm. Because there are no consistent approaches for evaluation of gene selection algorithms in the literature, it is not easy to compare the different algorithms using the published results alone. Therefore we employed the same technique, external B.632+, to assess the performance of different gene selection algorithms for comparison purpose in our study (the gene selection algorithms used for comparison are described below). The bootstrap samples are generated by resampling with replacement from the original dataset. In the present study, the balanced bootstrap samples with $K = 200$ replicates are employed to reduce the variance. Each sample in the original sample set is made to appear exactly K times in the balanced bootstrap samples. The SVM was employed as the classifier to estimate the error rates of different gene selection algorithms.

In the present study, we compared our LS Bound measure with a few feature selection criteria, including Fisher's ratio and Mahalanobis class separability measure. Fisher's ratio is an individual gene ranking criterion and is used to evaluate how well a single gene is correlated with the separation between classes. For every gene the Fisher's ratio is defined as $f = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$, where $\mu_1, \mu_2, \sigma_1, \sigma_2$ denote the means and standard deviations of two classes. The Mahalanobis class separability measure is a well-known feature subset evaluation criterion in the literature of pattern recognition (Devijver and Kittler, 1982). In the experiment, the Mahalanobis class separability measure is combined with SFS algorithm for gene selection. We also compared our gene selection algorithm with other published gene selection algorithms, including the Weighting factor employed both in the Weighted Voting algorithm (Golub *et al*., 1999) and with the SVM (Furey *et al*., 2000), and SVM RFE (Recursive Feature Elimination) (Guyon *et al*., 2002). The Weighting factor $[a = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)]$ is a minor variant of Fisher's ratio, and is also commonly used in the literature of microarray data analysis. Benefiting from the good performance of SVMs in high dimensional gene-expression data, SVM RFE is often considered as one of the best gene selection algorithm in the literature.

3.1 Colon cancer dataset

The colon cancer dataset (Alon *et al*., 1999) contains gene-expression levels of 40 tumor and 22 normal colon tissues for 2000 genes. The task is to identify important genes which can distinguish colon cancer from normal tissues.

We assessed the performance of our gene selection criterion using external .632+ bootstrap technique and compared it with Fisher's ratio and Mahalanobis class separability measure, and other published gene selection algorithms, such as Weighting factor and SVM RFE. As shown in Figure 1, LS Bound measure is slightly inferior to SVM RFE, but it outperforms other three algorithms. Note that only genes 1, 2, 4, 8, 16 and 32 are selected in SVM RFE due to its selection mechanism.

To identify the important genes for classification problem, we first identified 50 genes by employing the gene selection on the entire original dataset. According to the LS Bound measure on the selected gene subset, we finally selected the first 15 genes listed in Table 1, because adding more genes would not result in significant change on the LS Bound measure. Some comments about the selected genes are worthy of mention. The product of CD44 is a family of transmembrane glycoproteins generated by alternative splicing and differential glycosylation. One of the CD44 variant forms, CD44v6, is closely associated with transformation in human colon cancer, and its overexpression may be a clinical indicator of colon cancer (Yamada *et al*., 2003). Higher level of expression of RPS3 (Pogue-Geile *et al*., 1991) in colon cancer compared to normal tissue has also been observed before.

3.2 Leukemia dataset

The leukemia dataset was first described by Golub *et al*. (1999). The dataset contains gene-expression levels of 72 patients with either acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases) for 7129 human genes. The raw data are available at http://www-genome.wi.mit.edu/cancer/. The task is

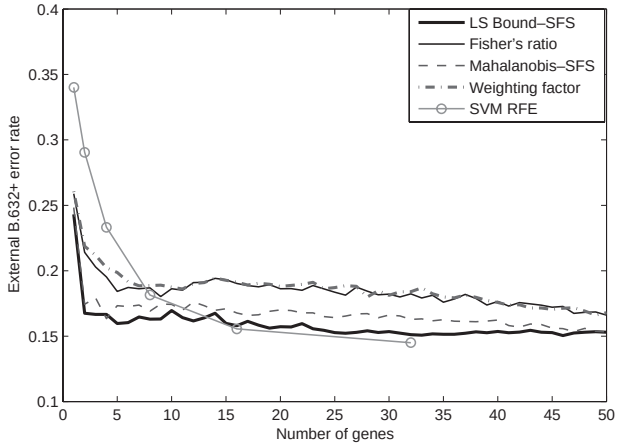

Fig. 1. The external B.632+ error for the colon cancer dataset, shown as the number of selected genes. The five curves are obtained from five gene selection algorithms: the LS Bound measure combining with SFS algorithm, Fisher's ratio, Mahalanobis class separability measure with SFS algorithm, Weighting factor and SVM RFE. The performance of LS Bound measure is quite good.

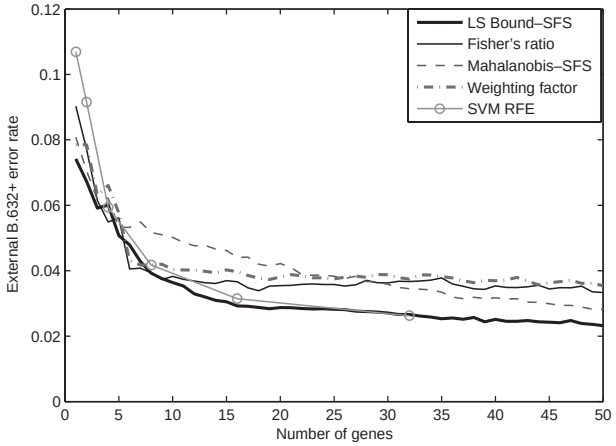

Fig. 2. The external B.632+ error for the leukemia dataset, shown as the number of selected genes. The five curves are obtained from five gene selection algorithms: the LS Bound measure combining with SFS algorithm, Fisher's ratio, Mahalanobis class separability measure with SFS algorithm, Weighting factor and SVM RFE. The LS Bound measure results in better performance than others.

Table 1. The top 15 selected genes for the colon cancer dataset

No.	Access no.	Gene
1	R87126	MYH, nonmuscle (*Gallus gallus*)
2	X55715	RPS3
3	T94579	Chitotriosidase precursor
4	T61661	PFN1
5	T57882	MYH9
6	R88740	ATP5J
7	T70062	ILF2
8	L37792	STX1A
9	M59042	CD44
10	K03474	MIS
11	T63508	Ferritin heavy chain
12	M84349	CD59
13	H15813	CEBPB
14	D00762	Proteasome comoponent C8
15	X53586	ITGA6

The genes and ESTs without indicating the source are all from *Homo sapiens*.

Table 2. The top 12 selected genes for the leukemia dataset

No.	Access no.	Gene
1	M84526	Adipsin
2	L15326	PTGS2(COX2)
3	M92287	Cyclin D3 (CCND3)
4	D26308	NADPH-flavin reductase
5	X95735	Zyxin
6	U75276	BRF1
7	U88667	ABCA4
8	U40343	CDKN2D
9	M31994	ALDH1
10	D26156	SMARCA4(SNF2)
11	J03779	CD10
12	S57212	MEF2C

All genes are from human mRNA.

to identify important genes which are discriminant between ALL and AML.

Again, we assessed the performance of our gene selection criterion using external .632+ bootstrap technique and compared it with Fisher's ratio, Mahalanobis class separability measure, Weighting factor and SVM RFE. As shown in Figure 2, our gene selection algorithm outperforms the other algorithms in this dataset.

By employing the gene selection on the entire dataset we selected 12 most important genes for the classification between ALL and AML. The 12 genes are shown in Table 2. Among the selected genes, some of them have been reported to be related to myeloid or lymphoblastic leukemia. Adipsin (D component of complement) is contained in the locus 19p13.3, whose chromatin reorganization is known to be associated with myeloid cell differentiation (Wong *et al*., 1999). Cyclin D3 encodes protein critical for controlling the physiological progression from the G_1 to the S phase of the cell cycle. In acute lymphoblastic leukemia (ALL) cells conditional expression of cyclin D3 leads to preventing glucocorticoid-induced cell cycle G_1 arrest (Ausserlechner *et al*., 2004). The gene ALDH1, which encodes the retinoic acid synthesizing enzyme, was found to be ectopically activated by HOX11 in NIH 3T3 cells, while the activation of HOX11 is a feature of some T-cell tumors and in many T-cell ALL cases the HOX11 gene is activated by translocation into either of two T-cell receptor loci (Greene *et al*., 1998). CD10, the common acute lymphoblastic leukemia antigen (CALLA), is the most common marker for clinical immunophenotypic classification of ALL. It is a membrane-bound neutral endopeptidase that can be expressed in both B- and T-cell ALL (Pui *et al*., 1993). In the work of Golub *et al*. (1999) four genes (adipsin, cyclin D3, zyxin and SMARCA4) are also selected. Although Zyxin gene, which encodes protein important for cell adhesion, and SMARCA4, which

encodes protein for chromatin remodeling, were not reported to have any role in hematopoisis, Zyxin is highly correlated with acute myelogenous leukemia (AML) while SMARCA4 is correlated with B-precursor ALL.

4 CONCLUSION

In the present study, we have proposed an LS Bound measure as the evaluate criterion for gene selection. The LS Bound measure can be considered as a hybrid of filter and wrapper methods. On the one hand, the LS Bound measure is derived from the leave-one-out procedure of LS-SVMs. As the upper bound for leave-one-out classification results, the LS Bound measure has direct relation to the performance of LS-SVMs, consequently it provides gene subset leading to more accurate classification results than the filter method. On the other hand, unlike the classical wrapper method, the LS Bound measure does not demand repeated trainings for cross validation. The training procedure involved is implicitly expressed in Equation (11), which has a similar computational complexity to the filter method. The effectiveness of the LS Bound measure has been tested on two benchmark microarray datasets when the LS Bound measure is combined with SFS to form a gene selection algorithm. We also combined the SFFS search procedure with LS Bound measure for gene selection on the same datasets, and the results revealed that the superiority of the LS Bound measure does not depend on a specific search procedure. For details, please refer to the website http://www.ntu.edu.sg/home5/pg02776030/lsbound/

REFERENCES

Alon,U., Barkai,N., Notterman,D., Gish,K., Ybarra,S., Mack,D. and Levine,A. (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.

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci.*, **99**, 6562–6566.

Ausserlechner,M.J., Obexer,P., Bock,G., Geley,S. and Kofler,R. (2004) Cyclin D3 and c-MYC control glucocorticoid-induced cell cycle arrest but not apoptosis in lymphoblastic leukemia cells. *Cell Death Differ.*, **11**, 165–174.

Braga-Neto,U.M. and Dougherty,E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Dash,M. and Liu,H. (1997) Feature selection for classification. *Intell. Data Anal.*, **1**, 131–156.

Devijver,P. and Kittler,J. (1982) *Pattern Recognition: a Statistical Approach*. Prentice Hall, London.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Efron,B. and Tibshirani,R. (1997) Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.

Furey,T., Cristianini,N., Duffy,N., Bednarski,D., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,G.H. and Van Loan,C.F. (1996) *Matrix Computations*, 3rd edn. The Johns Hopkins University Press, Baltimore, MD.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lander,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Greene,W.K., Bahn,S., Masson,N. and Rabbitts,T.H. (1998) The T-cell oncogenic protein HOX11 activates Aldh1 expression in NIH 3T3 cells but represses its expression in mouse spleen development. *Mol. Cell. Biol.*, **18**, 7030–7037.

Guyon,I., Weston,J., Barnhill,S. and Vapnik.V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Inza,I., Sierra,B., Blanco,R. and Larranaga,P. (2002) Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *J. Intell. Fuzzy Syst.*, **12**, 25–33.

Li,L., Weinberg,C.R., Darden,T.A. and Pedersen,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Luntz,A. and Brailovsky,V. (1969) On estimation of characters obtained in statistical procedure of recognition (in Russian). *Technicheskaya Kibernatica*, **3**.

Pogue-Geile,K., Geiser,J.R., Shu,M., Miller,C., Wool,I.G., Meisler,A.I. and Pipas,J.M. (1991) Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. *Mol. Cell. Biol.*, **11**, 3842–3849.

Pudil,P., Novovicova,J. and Kittler,J. (1994) Floating search methods in feature selection. *Pattern Recogn. Lett.*, **15**, 1119–1125.

Pui,C.H., Rivera,G.K., Hancock,M.L., Raimondi,S.C., Sandlund,J.T., Mahmoud,H.H., Ribeiro,R.C., Furman,W.L., Hurwitz,C.A. and Crist,W.M. (1993) Clinical significance of CD10 expression in childhood acute lymphoblastic leukemia. *Leukemia*, **7**, 35–40.

Suykens,J.A.K. and Vandewalle,J. (1999) Least squares support vector machine classifiers. *Neural Process. Lett.*, **9**, 293–300.

Vapnik,V. and Chapelle,O. (2000) Bounds on error expectation for support vector machines. *Neural Comput.*, **12**, 2013–2036.

Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York.

Wong,E.T., Jenne,D.E., Zimmer,M., Porter,S.D. and Gilks,C.B. (1999) Changes in chromatin organization at the neutrophil elastase locus associated with myeloid cell differentiation. *Blood*, **94**, 3730–3736.

Xiong,M., Fang,X. and Zhao,J. (2001) Biomarker identification by feature wrappers. *Genome Res.*, **11**, 1878–1887.

Yamada,Y., Itano,N., Narimatsu,H., Kudo,T., Hirohashi,S., Ochiai,A., Tohnai,I., Ueda,M. and Kimata,K. (2003) CD44 variant exon 6 expressions in colon cancer assessed by quantitative analysis using real time reverse transcriptase–polymerase chain reaction. *Oncol. Rep.*, **10**, 1919–1924.