Multidimensional BSS/WSS Criterion for Modified SBFS Feature Selection Method in Tumor Classification

Hongyi Peng and Chunfu Jiang

ABSTRACT

To avoid the defect of BSS/WSS criterion, we propose a multidimensional BSS/WSS feature selection criterion and modify the sequential backward floating selection (SBFS) algorithm to deal with the case where the covariance matrix is singular in this study. Then, we use support vector machine (SVM) to classify the gene expression data based on the proposed feature selection algorithm. The performance of the proposed approach is compared with BSS/WSS criterion and some other popular methods in feature selection and classification via the well-known colon cancer and prostate datasets in microarray literature, which demonstrates that the proposed criterion can take into account genes' joint discriminatory power, and the proposed feature selection method can obtain correct and informative gene subset for tumor classification.

KEYWORDS

Multidimensional BSS/WSS Criterion, Feature Selection, Modified SBFS Algorithm, Tumor Classification

Hongyi Peng, College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, PR China
Chunfu Jiang, College of Mathematics and Computational Science, Shenzhen University, Shenzhen 518060, P R China
Correspondence should be addressed to Hongyi Peng: penghyi@163.com
INTRODUCTION

The development of microarray technology increases the possibility of cancer classification and diagnosis at the gene expression level [1,2]. The problem of analysis of gene expression data is challenging because of the number of genes is usually much larger than the number of samples available. Jain et al.[3] points out that there are too many features, which may be irrelevant to the analysis actually, degrade the generalization performance of a classifier. Thus, identification of genes (i.e., features) that contribute most to the classification is critical to improve the accuracy and speed of prediction systems.

In the past decade, there are a vast amount of literatures reported, which focused on how to use a feature selection method for tumor classification [4,5,6,7]. In addition, several methods for reducing the number of genes to be considered before using the appropriate classification, are univariate methods in the sense that each relevant gene is considered individually. Examples include the weighted voting scheme, signal-to-noise (SNR) [4], the mixture model algorithm [8], the Wilcoxon test statistic [9], and the ratio of their between-group to within-group sums of squares (BSS/WSS) [5]. But BSS/WSS is single variable criterion and it cannot reflect variables’ joint discriminatory power.

To take into account the dependency between genes for achieving a reduced number of relevant genes, some multivariate gene selection procedures have been proposed. For example, generalized singular g-prior SSVS (gsg-SSVS, hereafter) [10], variable selection using the modified SFFS algorithm based on weighted Mahalanobis distance (MSWM, hereafter) [11], feature selection for support vector machine that achieve sparsity by augmenting the SVM objective with Lq-norm penalty term via difference of convex functions (DC) (DC-Lq-SVM, hereafter) [12], variable selection for Fisher linear discriminant analysis using the modified sequential backward selection algorithm [13]. The gsg-SSVS method performed posterior inference by concentrating on the posterior distribution and implement a stochastic search variable selection (SSVS) algorithm. Thus the feature selection method of gsg-SSVS doesn't succeed to obtain accurate genes' joint discriminatory power. For the difficulty of defining a proper DC composition of the objective function, the DC-Lq-SVM method [12] may fail to take into account genes’ joint discriminatory power. Optimal selection methods such as exhaustive search method are not practical for very high-dimensional problems. Therefore, alternative suboptimal methods, such as sequential forward floating and sequential backward floating selection methods (SFFS and SBFS) are in Pudil et al.[14] Compared with SFFS, SBFS doesn't omit significant features, it is more general. But SBFS method cannot deal with the case of a singular covariance matrix. However, when the number of samples is less than the number of genes as in many gene expression profiling studies, the matrix is likely to be singular, hence resulting in a numerical problem in calculating the inverse matrix. Therefore, we propose a multidimensional BSS/WSS (M_BW) criterion and modify the SBFS algorithm to deal with such
cases in this paper. We call our feature selection as M_BW in brief. We illustrate the advantage of our method on two well-known microarray datasets: colon cancer data [1] and prostate data [15] and use support vector machine (SVM) to classify the gene expression data.

The remainder of the paper is organized as follows. Section 2 proposes a multidimensional BSS/WSS feature selection criterion. Section 3 proposes a modified SBFS algorithm to deal with the case where the covariance matrix is singular. Section 4, an experiment is carried out to demonstrate the proposed M_BW method. Finally, the conclusion is provided in section 5.

MULTIDIMENSIONAL BSS/WSS CRITERION

In order to select the feature subset whose joint discriminatory power is maximum among all the subsets of the same number of features, the key is to consider the correlation between features and their joint discriminatory. The one-dimensional BSS/WSS criterion is defined as follows [5]

$$J(x) = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_k - \bar{x})^2}{\sum_i \sum_k I(y_i = k)(x_i - \bar{x}_k)^2}$$

Where $x$ is a column vector, denotes a gene, $x_i$ denotes the $i$-th sample expression level of gene $x$, $\bar{x}$ denotes the average expression level of gene $x$ and $\bar{x}_k$ denotes the average expression level of gene $x$ across samples belonging to class $k$.

We extend it to a multidimensional situation. The M_BW criterion can be defined in the followings:

$$J(X) = \frac{\sum_i I(y_i = k)(\bar{x}_k - \bar{x}) \Sigma_x^{-1} (\bar{x}_k - \bar{x})^t}{\sum_i I(y_i = k)(x_i - \bar{x}_k) \Sigma_x^{-1} (x_i - \bar{x}_k)^t}$$

Where $X$ denotes the data set of gene expression, where the number of rows of $X$ is equaled to the number of samples, the number of rows of $X$ is equal to the number of genes (features) being selected, $x_i$ is the $i$-th row vector of $X$, $\Sigma_X$ is the covariance matrix of $X$, and $\bar{x}_k$ denotes the mean vector of class $k$. Accordingly, we rank genes by the one-dimensional BSS/WSS criterion, and select the top $m$ genes as the initial gene subset.
MODIFIED SBFS FEATURE METHOD

When the number of samples is less than the number of genes as in many gene expression profiling studies, the covariance matrix $\Sigma_X$ is likely to be singular, hence resulting in a numerical problem in calculating $\Sigma_X^{-1}$. But the SBFS method [14] cannot deal with the case. In order to solve this problem, we propose a modified SBFS method to deal with the case.

Modified SBFS Procedure

We can design the modified SBFS algorithm by the following four steps:

Step 1. Start with the complete set $X_0 = X$.

Step 2. If the matrix $\Sigma_X$ is singular, then compute the M_BW according to section 3.2 and delete feature $x_{k+1}$ from the set $X_k$, to form feature set $X_{k+1}$, i.e. the least significant feature $x_{k+1}$ is deleted from the set $X_k$. Therefore $X_{k+1} = X_k - x_{k+1}$, and go to step 2. Else if the matrix $\Sigma_X$ is not singular, then go to step 3.

Testing singularity using $|\det (\Sigma_X)| \leq \text{tolerance}$ is not recommended as it is difficult to choose the correct tolerance. The condition number of $\Sigma_X$ can check for singular and nearly singular matrices. For example, if the condition number of $\Sigma_X$ is greater than 1000, then the matrix $\Sigma_X$ is singular.

Step 3. According SBFS algorithm to search feature subsets in data set $X_k$, where every feature subset has the largest multidimensional BW if the numbers of features are the same.

Suppose $k$ features have already been removed from the complete set of measurements $F_0 = X_k$ to form feature set $F_k$ with the corresponding criterion function $J(F_k)$. Furthermore, the values of all superset $F_i, i = 1, 2, \ldots, k - 1$, are known and stored, where $X_k$ has $E$ available features.

Step 4. Compute the prediction accuracy $P(F_k), k = 1, 2, \ldots, E$, base on k-fold cross-validation and obtain the optimal feature subset $F_k$ with the best predict accuracy. If there are several feature subsets have the same best predict accuracy, then we select the $F_k$ with the least number of features as the optimal feature subset.
Algorithm for Computing $M_{BW}$ in Singular Situation

Suppose the threshold value of condition number is $C$. If the condition number of matrix $\Sigma_x$ is greater than $C$, then the matrix is singular, and we can compute the $J(X)$ as follows:

Step 1. Perform principal component analysis on the matrix $\Sigma_x$, i.e. $[V,L]=P(\Sigma_x)$, where $P(*)$ denotes a function which performs the principal components analysis, $V$ is a $p$-by-$p$ matrix, with each column containing coefficients for one principal component. The columns are in order of decreasing component variance. $L$ is a vector containing the principal component variances, that is, the eigenvalues of $\Sigma_x$ are in order of decreasing component variance.

Step 2. Let $p_0 = \min \{ i | L(1,1)/L(i,1) > C \}$.

Step 3. Let $L2(i,1) = 1/L(i,1), 1 \leq i < p, L2(i,1) = 0, i \geq p_0$, where $L2$ is a column, and it has the same number of rows with the vector $L$.

Step 4. Let $J(X) = \sum_k I(y_i = k)(\overline{X_i} - \overline{X}) V \cdot D(L2) \cdot V'(\overline{X_i} - \overline{X})$, where $D(L2)$ returns a square matrix, with the elements of $L2$ on the main diagonal.

EXPERIMENTAL RESULTS

We illustrate the usefulness of the proposed $M_{BW}$ approach via two well-known datasets: the colon cancer data analyzed initially by Alon et al. [1] and the prostate cancer data analyzed by Singh et al. [15].

Colon Dataset

The colon dataset contains 62 samples of which 22 are from normal colon tissues and the remaining from colon cancer[1], each including gene expression values for 2000 different genes. In this experiment, we rank the discriminatory power of an individual gene according to the one-dimensional BSS/WSS criterion firstly, and obtain 150 top ranked genes. The threshold value of the condition number is 500 in this experiment. In the second step of the modified SBFS procedure, we choose out 29 genes from the 150 genes based on $M_{BW}$ criterion. In the fourth step of modified SBFS procedure, we compute the mean misclassification rates (MCRs) of 29 marked feature subsets and Figure 1 is the stem plot of the MCRs of 29 marked feature subsets. From figure 1, we can see that the feature subset with 10 features have the least MCR. We selected 10 genes, which are
H20709 (14), T57619 (43), T51529 (127), M76378 (267), D26129 (350), R87126 (493), M76378 (765), R88740 (792), X74295 (1247) and R54097 (1473), where the numbers in the parentheses are the indices of features in the original colon cancer dataset. We find that the M76378 gene has three indices in the colon cancer dataset, which are 245, 267,765. We choose out two indices of M76378, i.e. 267 and 765.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of genes</th>
<th>LOOCV error rate</th>
<th>10-fold CV MCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM [7]</td>
<td>1000 or 2000</td>
<td>9.68%</td>
<td>10.12%</td>
</tr>
<tr>
<td>LogitBoost, 100 iterations [6]</td>
<td>10</td>
<td>14.52%</td>
<td>15.22%</td>
</tr>
<tr>
<td>AdaBoost, 100 iterations [6]</td>
<td>10</td>
<td>16.13%</td>
<td>17.13%</td>
</tr>
<tr>
<td>gsg-SSVS [10]</td>
<td>14</td>
<td>11.29%</td>
<td>11.93%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11.29%</td>
<td>11.93%</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>11.29%</td>
<td>11.93%</td>
</tr>
<tr>
<td>DC-Lq-SVM</td>
<td>10</td>
<td>16.71%</td>
<td>17.92%</td>
</tr>
<tr>
<td>SRC-LatLRR</td>
<td>1000</td>
<td>7.82%</td>
<td>8.06%</td>
</tr>
<tr>
<td>BSS/WSS-SVM</td>
<td>5</td>
<td>8.06%</td>
<td>9.03%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11.29%</td>
<td>11.58%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>16.13%</td>
<td>16.35%</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>20.97%</td>
<td>20.94%</td>
</tr>
<tr>
<td>MSWM-SVM</td>
<td>7</td>
<td>8.06%</td>
<td>8.99%</td>
</tr>
<tr>
<td>M_BW-SVM</td>
<td>10</td>
<td>4.84%</td>
<td>5.07%</td>
</tr>
</tbody>
</table>

Figure 1. The stem plot of the MCRs of 29 marked feature subsets for colon dataset.

It is common to evaluate the performance of the classification methods for a selected subset of features by the LOOCV and 10-fold CV procedures. The summary of performance is presented in TABLE I. The M_BW method selected 10 features. When using the LOOCV procedure, our method misclassified one tumor tissue (T30) and two normal tissues (N34, N36). The summary is presented in TABLE I. It is clear from the comparison that our method is much better than the other popular classification methods.

From TABLE I, when using the 10-fold CV procedure, we can also see that our proposed classification method outperform the other classification methods, i.e.,
DC-Lq-SVM [12], SRC-LatLRR[16] and MSWM- SVM [11]. At the same time, we choose \( P = 5, 10, 20, 40 \) genes identified by the one-dimensional BSS/WSS criterion, and their performance are all worse than M_BW criterion’s. Therefore, we think that the M_BW criterion is reasonable for measuring genes' joint discriminatory power, and our feature selection method can obtain a good informative feature subset on the colon tumor classification.

Furthermore, some of the genes selected by our method are known to be associated with the Colon cancer of leukemia cells. For example, CRP (M76378) gene polymorphism is associated with increased colon cancer risk in African Americans compared to whites in the North Carolina colon cancer study [17]. At the same time, we find that the J02854 and M63391 genes rank 2 and 3 of the top 150 genes, respectively, which suggests their individual gene discriminatory powers are very strong. However, such genes are not included in our analysis. Nevertheless, we think that if J02854 or M63391 gene is combined with other genes, their joint discriminatory power cannot be the best.

**Prostate Dataset**

This dataset provides the expression levels of 12600 genes for 50 normal tissues and 52 prostate cancer tissues [15]. The experiment was run on Affymetrix human 95aV2 arrays. Oligonucleotide microarrays contain probes for approximately 12,600 genes and ETSs. In this experiment, we rank the discriminatory power of an individual gene according to the one-dimensional BSS/WSS firstly, and obtain 200 top ranked genes. The threshold value of the condition number is 500 in this experiment. In the second step of the modified SBFS procedure, we choose out 19 genes from the 200 genes M_BW criterion. In the fourth step of modified SBFS procedure, we compute the MCRs of 19 marked feature subsets and Figure 2 is the stem plot of the MCRs of 19 marked feature subset. From figure 2, we can see that the feature subset with 18 features have the least MCR. We selected 8 genes, which are AF037643 (996), M30894 (4365), T89651 (7304), Z11692 (8552), Z48199 (9133), AL036744 (10138), AF055376 (10234) and U21689 (11871), where the numbers in the parentheses are the indices of features in the original prostate cancer dataset.

![Figure 2. The stem plot of the MCRs of 19 marked feature subsets for prostate dataset.](image-url)
<table>
<thead>
<tr>
<th>Method</th>
<th>Number of genes</th>
<th>LOOCV error rate</th>
<th>10-fold CV MCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT-PLSDA [18]</td>
<td>18</td>
<td>8.82%</td>
<td>9.08%</td>
</tr>
<tr>
<td>DC-Lq-SVM</td>
<td>12</td>
<td>8.6%</td>
<td>8.80%</td>
</tr>
<tr>
<td>SRC-LatLRR</td>
<td>1000</td>
<td>3.54%</td>
<td>3.68%</td>
</tr>
<tr>
<td>BSS/WSS-SVM</td>
<td>5</td>
<td>6.86%</td>
<td>7.29%</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>7.84%</td>
<td>8.49%</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>8.82%</td>
<td>8.76%</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>9.80%</td>
<td>10.24%</td>
</tr>
<tr>
<td>MSWM-SVM</td>
<td>11</td>
<td>2.94%</td>
<td>3.26%</td>
</tr>
<tr>
<td>M_BW-SVM</td>
<td>8</td>
<td>1.96%</td>
<td>2.42%</td>
</tr>
</tbody>
</table>

As the same as colon cancer data section, we evaluate the performance of the classification methods for a selected subset of features by the LOOCV and 10-fold CV procedures for prostate cancer dataset. The M_BW method selected 8 features. The summary is presented in TABLE II. It is clear from the comparison that our method is much better than the other popular classification methods.

From TABLE II, when using the 10-fold CV procedure, we can also see that our proposed classification method outperform the other classification methods, i.e., DC-Lq-SVM [12], SRC-LatLRR [16] and MSWM- SVM [11]. At the same time, we chose P = 5, 8, 15, 30 genes identified by the one-dimensional BSS/WSS criterion, and their performance are all worse than the M_BW criterion’s. Therefore, we think that the M_BW criterion is reasonable for measuring genes' joint discriminatory power, and our feature selection method can obtain a good informative feature subset on the prostate tumor classification.

Furthermore, some of the genes selected by our method are known to be associated with the prostate cancer. For example, antibody induction to TARP (M30894) may represent a possible biomarker for treatment response to GM-CSF secreting cellular immunotherapy in prostate cancer patients [19], HEPSIN (X07732) gene is one of the most consistently overexpressed genes in patients with prostate cancer [20]. At the same time, we find that the AF035283 and AL050152 genes rank 3 and 4 of the top 200 genes, respectively, which suggests their individual gene discriminatory powers are very strong. However, such genes are not included in our analysis. Nevertheless, we think that if AF035283 or AL050152 gene is combined with other genes, their joint discriminatory power cannot be the best.

According to the analysis above, we can summary that for the classification of the colon and prostate cancer data sets, the M_BW criterion is reasonable for measuring genes' joint discriminatory power and the M_BW method can obtain a good informative gene subset.
CONCLUSIONS

A reliable and precise classification of tumors is essential for successful treatment of cancer. Microarray technology increases the possibility of cancer classification and diagnosis at the gene expression level. However, many factors may affect the outcome of the analysis. One of them is the huge number of genes included in the original data. Some of them may be irrelevant to the analysis. Thus, selecting informative genes is critical to improving the accuracy and speed of prediction systems. At the same time, a reasonable criterion is very important to feature selection.

In this paper, we propose a M_BW criterion to avoid the defect of BSS/WSS criterion and propose a modified SBFS method to deal with the case where the covariance matrix is singular. We apply our M_BW method to two different gene expression data sets: the colon and prostate datasets. We have compared the M_BW method with other popular methods, for example, LogitBoost, AdaBoost, gsg-SSVS, MSWM, RT-PLSDA, SRC-LatLRR, DC-Lq-SVM and BSS/WSS methods. The results show that the M_BW criterion is reasonable for measuring genes' joint discriminatory power and the M_BW method can obtain a good informative gene subset for tumor classification.

ACKNOWLEDGEMENT

This work is supported by the National Statistical Science Research Project of China under No. 2015LY81 and No.2016LY98.

REFERENCES