Structured polychotomous machine diagnosis of multiple cancer types using gene expression

Ja-Yong Koo¹,*, Insuk Sohn¹, Sujong Kim²,³ and Jae Won Lee¹

¹Department of Statistics, Korea University, Seoul 136-701, Korea, ²Department of Biochemistry, College of Medicine, Hanyang University, Seoul 133-791, Korea and ³Current address: Skin Research Institute, AmorePacific R&D Center, Yongin 449-729, Korea

Received on December 21, 2005; revised and accepted on January 26, 2006
Advance Access publication February 1, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: The problem of class prediction has received a tremendous amount of attention in the literature recently. In the context of DNA microarrays, where the task is to classify and predict the diagnostic category of a sample on the basis of its gene expression profile, a problem of particular importance is the diagnosis of cancer type based on microarray data. One method of classification which has been very successful in cancer diagnosis is the support vector machine (SVM). The latter has been shown (through simulations) to be superior in comparison with other methods, such as classical discriminant analysis, however, SVM suffers from the drawback that the solution is implicit and therefore is difficult to interpret. In order to remedy this difficulty, an analysis of variance decomposition using structured kernels is proposed and is referred to as the structured polychotomous machine. This technique utilizes Newton–Raphson to find estimates of coefficients followed by the Rao and Wald tests, respectively, for addition and deletion of import vectors.

Results: The proposed method is applied to microarray data and simulation data. The major breakthrough of our method is efficiency in that only a minimal number of genes that accurately predict the classes are selected. It has been verified that the selected genes serve as legitimate markers for cancer classification from a biological point of view.

Availability: All source codes used are available on request from the authors.

Contact: jykoo@korea.ac.kr

1 INTRODUCTION

DNA microarray analysis is a new biotechnological breakthrough, which allows the simultaneous monitoring of thousands of gene expressions in cells (Brown and Botstein, 1999) and has far reaching applications in pharmaceutical and clinical research. By comparing gene expression in normal and tumor tissues, for example, microarrays can be used to identify tumor-related genes and targets for therapeutic drugs (Alizadeh et al., 2000).

Classification of different phenotypes, predominantly cancer types, using microarray gene expression data is abundant in the literature; see e.g. Golub et al. (1999), Alizadeh et al. (2000), Furey et al. (2000) and Dudoit et al. (2002). The methods used in these studies range from classical discriminant analysis to flexible tools from machine learning such as bagging, boosting and support vector machines (SVM). SVM has been successfully applied to classification of gene expression data; Furey et al. (2000), Ramaswamy et al. (2001) and Guyon et al. (2002) used SVM and Lee and Lee (2003) adopted multiclass SVM (Lee et al., 2004) for this purpose. Recently, Lee et al. (2005) compared the performance of various classification methods, and provided guidelines for the most appropriate classification tool in various contexts.

SVM is growing in popularity as a classification problem tool with abundantly many successful and diverse applications (Vapnik, 1998; Schölkopf et al., 2002). Recently, Zhu and Hastie (2001) proposed replacing the hinge loss by the multinomial likelihood coined this method import vector machine (IVM). It can be shown that IVM has advantages over SVM in the following aspects:

1. IVM can naturally handle the polychotomous classification problem;
2. IVM can provide estimates of the posterior probabilities and
3. the computational cost of the IVM is typically cheaper than that of SVM (Zhu and Hastie, 2001).

Ordinary SVM suffers from the drawback that the solution is implicit and therefore is difficult to interpret. A method for improving the interpretability of SVM is the SVM-RFE method which has been introduced by Guyon et al. (2002). In this paper, we propose structured polychotomous machine (SPM) which extends IVM through a functional analysis of variance (ANOVA) decomposition using structured kernels. We use Newton–Raphson to find the coefficient estimates followed by the Rao and Wald tests, respectively, for addition and deletion of import vectors (Zhu and Hastie, 2001). A computational improvement over IVM is the usage of the Rao statistic which can be calculated faster than the one-step Newton–Raphson method of Zhu and Hastie (2001). The Wald statistic for the stepwise deletion step can be carried out with an insignificant amount of extra computation, because the deletion algorithm is much less computer intensive than the addition algorithm.
A summary of this paper is as follows. Section 2 describes the SPM and the stepwise algorithm for the selection of the import vectors. Section 3 illustrates the performance of the proposed algorithm using real microarray data and simulated data as examples. Concluding remarks and discussions are given in Section 4.

2 SYSTEMS AND METHODS

2.1 Structured polychotomous machines

In this section, we propose the SPM based on a functional ANOVA decomposition using structured polynomials.

Consider the training data \( L = \{(x_n,y_n) : n = 1, \ldots, N\} \), where the input \( x_n \) belongs to some domain \( X \subseteq \mathbb{R}^d \) and the label \( y_n \in M = \{0, 1, \ldots, M\} \). Let the input space be given by some subsets \( A \subseteq \mathbb{R}^d \) such that the function \( h \) is a constant, \( h \) are the main effects, \( h_{ij} \) are the two-factor interactions, and so on. For example, additive models have the form of \( h_n = b + h_1(x_1) + \cdots + h_d(x_d) \).

We will consider the additive SPM alone because the main focus is on the feature (gene) selection. The SPM methodology can handle two-factor interactions. For \( 1 \leq j \leq d \), let \( n_j \) be a non-negative integer and \( S' = \{s'_i : i = 1, \ldots, n_j\} \subseteq \mathcal{P}' = \{x_1, \ldots, x_d\} \) be the set of import vectors for the \( j \)-th coordinate. Let \( S = \bigcup_{i=1}^{n_j} S' \) and \( \mathcal{P} = \bigcup_{i=1}^{n} \mathcal{P}' \). We refer to \( S \) as the set of import vectors following Zhu and Hastie (2001). Suppose the kernel function for each coordinate are equal, i.e. \( K = 1 \leq j \leq d \). Let \( J = 1, \ldots, n_j \) and denote by \( B_1, \ldots, B_J \) the basis functions consisting of the constant basis function 1 and the univariate functions \( K(s_i) \), where \( s_i \in S \). Consider

\[
\eta_j(m|x; \beta_m) = \beta_{m1} B_1(x) + \cdots + \beta_{mJ} B_J(x),
\]

where the additive regularization matrix corresponding to \( S' \) is given by

\[
K' = \begin{pmatrix}
K(s'_1, s'_1) & \ldots & K(s'_1, s'_J) \\
\vdots & \ddots & \vdots \\
K(s'_1, s'_J) & \ldots & K(s'_J, s'_J)
\end{pmatrix}
\]

It can be noted from (3) that the constant basis function is not regularized. The (regularized) maximum likelihood estimator (MLE) \( \hat{\beta} \) is as the maximizer of \( \ell_j(\beta) \). The SPM classifies a new input \( x \) into the class \( m \) which is defined by

\[
m = \arg \max_{m \in M} p(m|x; \beta_m).
\]

In order to find \( \hat{\beta} \), we use the Newton–Raphson method. Let \( S_j(\beta) = \partial \ell_j(\beta)/\partial \beta \) denote the score at \( \beta \) and let \( \ell_j(\beta) \) denote the information matrix with entries \( -\partial^2 \ell_j(\beta)/\partial \beta^2 \). The maximum likelihood estimator \( \hat{\beta} \) satisfies the likelihood equations \( S_j(\beta) = 0 \). The Newton–Raphson method for computing \( \hat{\beta} \) is to iteratively determine \( \beta^{(n+1)} \) from \( \beta^n \) according to the formula

\[
\beta^{(n+1)} = \beta^n + S_j(\beta^n)^{-1} S_j(\beta^n).
\]

2.2 The Rao statistic and the Wald statistic

In this section, we explain the Rao (score) statistic and the Wald statistic (Rao, 1973) which will be used in choosing import vectors. The dependency of \( \hat{\beta} \) on \( \lambda \) is suppressed for notational convenience.

2.2.1 Rao statistic

Given \( B_1, \ldots, B_J \), let \( \hat{\beta} = (\hat{\beta}_m) \) and \( \ell_j(\beta) \) be the MLE and the information matrix, respectively. Here, \( \beta_m \) denotes the estimated coefficient corresponding to the \( j \)-th class and the \( k \)-th basis function \( B_k \). Let \( B_{1j} \) be a candidate basis function for addition. Define \( \tilde{\beta} = (\tilde{\beta}_m) \) by

\[
\tilde{\beta}_m = \begin{cases} 
\hat{\beta}_m & 1 \leq j \leq M, 1 \leq k \leq J \\
0 & 0 \leq j \leq M, k = J + 1.
\end{cases}
\]

Then the Rao statistic is defined by

\[
R = \frac{S_j(\beta^*) - S_j(\hat{\beta})}{\text{det}(S_j(\hat{\beta}))}.
\]

2.2.2 Wald statistic

Given \( B_1, \ldots, B_J \), let \( \tilde{\beta} = (\tilde{\beta}_m) \) and \( \ell_j(\beta) \) be the MLE and the information matrix, respectively. Let \( B_k \) be the candidate basis function for deletion and \( \tilde{\beta} \) denote \( M \times 1 \) vector of elements \( \tilde{\beta}_{mk}, m = 1, \ldots, M \). The Wald statistic is defined by

\[
W = \tilde{\beta}^T [\text{cov}(\tilde{\beta})]^{-1} \tilde{\beta},
\]

where \( \text{cov}(\tilde{\beta}) \) is the \( M \times M \) submatrix of \( [\ell_j(\beta)] \) whose rows and columns correspond to these \( M \) coefficients. The Wald statistic can be used as a measure of the importance of covariates which are selected in the final SPM.
In order to find the final model, one needs a model selection criterion. During the combination of stepwise addition and deletion, we get a sequence of models indexed by $v$ with the $v$-th model having $M_J$, parameters and MLE $\hat{\beta}_2$. We select the model that minimizes

$$\text{BIC}_v = 2\ell(v, \hat{\beta}_2) + (\log N) \cdot (M_J),$$

which is similar to the BIC criterion of Schwarz (1978). One may also use the test error rate as the model selection criterion for the final SPM model. The results of this paper have been obtained using BIC.

**ANOVA Sparse Stepwise Algorithm (ASSA)**

Require: L and P
Fit the constant model with $\lambda = 2^9$ and $\hat{S} = \emptyset$
Reduce $\lambda$ by two

Repeat addition with given $\lambda$
Pick $\hat{s}^i = \text{argmax}_{s \subseteq \hat{S}} R(s)$, $1 \leq i \leq d$.
Find $j^* = \text{argmax}_{j \in \hat{S}} R(\hat{s}^i)$.
Define $S' = S' \cup \{\hat{s}^i\}$.
Compute $\hat{\beta}_1$ and test error rate.
Until $\ell(\hat{\beta}_1)$ does not increase much

Repeat deletion with optimal $\lambda$
Pick $\hat{s}^i = \text{argmin}_{s \subseteq \hat{S}} W(s)$, $1 \leq i \leq d$.
Find $j^* = \text{argmin}_{j \in \hat{S}} W(\hat{s}^i)$.
Define $S'' = S'' \setminus \{\hat{s}^i\}$.
Compute $\hat{\beta}_1$ and test error rate.
Until $\hat{S} = \emptyset$
Output: $S = \cup_{i=1}^{l} S'_i$ and $\hat{\beta}_1$ corresponding to the model having minimum BIC

### 3 RESULTS

This section presents the selection performance of SPM using real microarray data and simulated data. In order to show the gene selection performance, we adopt the additive SPM defined by (1) and (2) with a Gaussian kernel. Consider the problem of choosing $\sigma$ which is the tuning parameter of the gaussian kernel. It has been observed that the performance of SPM is not much sensitive to the choice of $\sigma$. In order to make SPM data-dependent, we adopt cross-validation for the choice of $\sigma$. Given $\sigma$, let $\text{SPM}_\sigma$ denote the SPM fitted to a training dataset according to the ASSA algorithm using BIC as the model selection criterion. The tuning parameter $\sigma$ were chosen by search over $2^0, \ldots, 2^9$ using 5-fold cross-validation. Given a training dataset, the SPM fit is defined by $\text{SPM}_\sigma$, where $\sigma$ is the minimizer of the cross-validated prediction error.

### 3.1 Simulation

We carried out a simulation study to evaluate SPM. We simulated artificial dataset assuming normal distributions of log expression levels as in Broberg (2002). The means and standard deviations for the simulation data are shown in Table 1. In Table 1, the first three rows represent the insignificant genes (null case), whereas the last three rows represent the significantly differentially expressed genes (significant case). Means and standard deviations were chosen randomly among the three rows in both null and significant cases. The sample sizes were (47 and 25) arrays, which are the same as those for the real leukemia dataset. Simulated data contained 1% significantly differentially expressed genes out of 1000 genes.

### 3.2 Real data

#### 3.2.1 Small round blue cell tumor

This dataset came from small round blue cell tumors (SRBCT) study (Khan et al., 2001). The data, consisting of expression values of 2308 genes, were obtained from cDNA microarrays, which were made according to the standard protocol of National Human Genome Research Institute. The SRBCTs were categorized to Burkitt lymphoma (BL): 8, Ewing sarcoma (EWS): 23, neuroblastoma (NB): 12 or rhabdomyosarcoma (RMS): 20. The dataset consisted of 63 training samples and 20 test samples. Logarithm base 10 of the expression levels was taken and the arrays were standardized. SPM selected 6 genes (Table 2) from the 63 training samples. Both the training and test error rates were 0%.

### Table 1. The means and standard deviations for the simulation data

<table>
<thead>
<tr>
<th></th>
<th>Mean1</th>
<th>SD1</th>
<th>Mean2</th>
<th>SD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>-8</td>
<td>0.2</td>
<td>-8</td>
<td>0.2</td>
</tr>
<tr>
<td>case</td>
<td>-10</td>
<td>0.4</td>
<td>-10</td>
<td>0.4</td>
</tr>
<tr>
<td>Significant</td>
<td>-12</td>
<td>1</td>
<td>-12</td>
<td>1</td>
</tr>
<tr>
<td>case</td>
<td>-6</td>
<td>0.1</td>
<td>-6.1</td>
<td>0.1</td>
</tr>
<tr>
<td>-8</td>
<td>0.2</td>
<td></td>
<td>-8.5</td>
<td>0.2</td>
</tr>
<tr>
<td>-10</td>
<td>0.4</td>
<td></td>
<td>-11</td>
<td>0.7</td>
</tr>
</tbody>
</table>

The first three rows represent the insignificant genes (null case), whereas the last three rows represent the significantly differentially expressed genes (significant case).

### Table 2. The average number of genes selected and the average number of true genes selected from the simulated data

<table>
<thead>
<tr>
<th>Methods</th>
<th>SPM</th>
<th>PAM</th>
<th>SVM-RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of genes selected</td>
<td>3.8</td>
<td>9.52</td>
<td>4.42</td>
</tr>
<tr>
<td>Average number of true genes selected</td>
<td>3.34</td>
<td>7.76</td>
<td>3.6</td>
</tr>
<tr>
<td>Recovery rate (%)</td>
<td>87.9</td>
<td>81.5</td>
<td>81.4</td>
</tr>
</tbody>
</table>

The recovery rate is defined as the percentage of the number of true significant genes out of the selected genes.

A simulated dataset was randomly divided into a training sample (67%) and a test sample (33%), where the training sample was used to fit SPM and the test error rates were computed by the test sample. This procedure was repeated 50 times. We compared the average number of genes selected, along with the average number of true genes selected from the simulated data and misclassification rate by our method and other two well-known methods such as PAM (Tibshirani et al., 2002) and SVM-RFE (Guyon et al., 2002). We used PAM method by choosing the optimal amount of shrinkage which minimizes test error rates under the restriction that the maximum number of selected genes <100 and we used SVM-RFE method, where the optimal gene set was chosen by minimizing the leave-one-out cross-validation error rate.

Table 2 shows the average number of genes selected, along with the average number of true genes selected from the simulated data. As shown in Table 2, SPM selected smaller average number of genes selected with higher recovery rate. We also display the boxplot of the misclassification rates on the simulated data (Fig. 1), and SPM gave lower misclassification rate.
which means the SPM method can predict the tumor class for both seen and unseen samples using the six genes with 100% accuracy. The Wald statistic present the relevance of variables.

We compare the number of genes in the optimal set selected by our method with those of PAM and SVM-RFE (Table 3). The result for artificial neural networks (ANNs) was taken from Khan et al. (2001). We used the multi-class SVM-RFE in the rfe package for the multiclass case. Multi-class SVM-RFE is achieved using the one-against-one approach and each two-class SVM classifier is described by a weight vector. The $k$-class classifier based on the one-against-one approach is characterized by $k(k-1)/2$ weight vectors, where $k = M + 1$. The sum of the absolute value of the weight vector coordinate is used to characterize the discriminant power of the associated feature. We used multi-class SVM-RFE method with 200 genes because of limitations in computer memory. Table 3 shows test error rates and number of genes selected for SRBCT data. SPM selected the smallest number of genes that can accurately classify samples.

We also examined the correspondence of genes selected by SPM with ANNs, PAM and SVM-RFE methods. We found that SPM was high correspondence between all methods. All six genes selected by SPM were also selected by ANNs and four genes by PAM and SVM-RFE (Table 4).

Whether the selected genes serve as legitimate markers for cancer classification was further verified by cluster analysis and visualization. In this regard, we applied a hierarchical clustering program developed by Eisen (Eisen et al., 1998) to the expression data of the selected genes and then visualized the structure of the data (Fig. 2). Figure 3 illustrates six different patterns of the chosen six genes in the same fashion as Figure 2 in Lee and Lee (2003).

The SPM classifier was applied to the original data consisting of 2308 genes. It was the objective to detect important genes for classifying SRBCTs common in children, and at the same time to gauge the uncertainty that inherently lies in estimating the effects of the 2308 response covariates with only 63 observations. For the assessment of variability, 100 bootstrap samples are drawn from the training data with the same class proportions as in the original sample. Similar to Figure 7 in Lee et al. (2004), our Figure 4 shows the proportion of selecting each gene in 100 replicated SPM classifiers based on bootstrap samples in which four genes out of the selected six genes are consistently selected >45% of the time.

Previous studies already demonstrated that the six genes were predictive for subtype classification of SRBCT (Khan et al., 2001; Tibshirani et al., 2002). In addition, the relevance of some genes to specific types of tumor was reported in the biological literature. For example, over-expression of cyclin D1 in EWS and NB has been shown by biological methods (Zhang et al., 2004; Dauphinot et al., 2001; Molenaar et al., 2003; Elenitoba-Johnson et al., 2002). Insulin-like growth factor 2 (somatomedin A: IGF2), related to myogenesis, was also previously reported to be highly expressed in RMS (El-Badry et al., 1990; Khan et al., 1999). However, IGF2 is expressed in some other cancers and normal tissues, lacking specificity. Some genes that are under-represented in a particular type of tumor compared with other types can also be selected as predictive genes. For instance, cyclin-dependent kinase 6 (CDK6) gene selected for EWS was under-expressed in this tumor. Previously, CDK6 was reported to restrain proliferation in a certain type of cell and its loss or down-regulation was implicated to play a role in development and progress of some tumor types (Lucas et al., 2004). However, CDK6 is ubiquitously expressed in a wide variety of cells and strong expression was described in specific tumors such as acute lymphoblastic leukemia (Fink and LeBein, 2001; Chilosi et al., 1998; Omura-Minamisawa et al., 2000). This gene was identified in Khan et al.’s work and Tibshirani et al.’s work for SRBCT.

Meningioma (disrupted in balanced translocation) 1 (MNI1) was

![Fig. 1. Boxplots of the misclassification rates on the simulated data.](https://academic.oup.com/bioinformatics/article-abstract/22/8/950/227018)

**Table 3.** Test error rates and number of genes selected for SRBCT data

<table>
<thead>
<tr>
<th>Methods</th>
<th>ANNs</th>
<th>PAM</th>
<th>SPM</th>
<th>SVM-RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test error rates (%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>No. of genes</td>
<td>96</td>
<td>43</td>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table 4.** The gene list selected for the SRBCT data

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Wald statistic</th>
<th>ANNs</th>
<th>PAM</th>
<th>SVM-RFE</th>
<th>BSS/WSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin-like growth factor 2 (somatomedin A)</td>
<td>5.54</td>
<td>O</td>
<td>O</td>
<td>×</td>
<td>12</td>
</tr>
<tr>
<td>Cyclin D1 (PRAD1: parathyroid adenomatosis 1)</td>
<td>5.00</td>
<td>O</td>
<td>×</td>
<td>×</td>
<td>44</td>
</tr>
<tr>
<td>Fibronectin type III domain containing 5 (FNDC5)</td>
<td>4.68</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>16</td>
</tr>
<tr>
<td>Meningioma (disrupted in balanced translocation) 1</td>
<td>4.61</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>38</td>
</tr>
<tr>
<td>Cyclin-dependent kinase 6 (CDK6)</td>
<td>4.56</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>5</td>
</tr>
<tr>
<td>Major histocompatibility complex, class II, DP beta 1</td>
<td>4.31</td>
<td>O</td>
<td>×</td>
<td>O</td>
<td>116</td>
</tr>
</tbody>
</table>

The abbreviations used is: the ratio of between classes sum of squares to within class sum of squares (BSS/WSS). ‘O’ denotes the genes selected by ANNs, PAM and SVM-RFE methods. BSS/WSS column present rank.
also under-expressed in NB subtype. MN1 gene resides on chromosome 22 and was found to be disrupted by a balanced translocation (4;22) in meningioma, common benign brain tumor (Lekanne et al., 1995). Absence of functional MN1 protein was suggested to contribute to meningioma pathogenesis. In addition, MN1 was shown to be fused to TEL, a member of the family of ETS transcription factor on chromosome 12p13 (12;22) in acute myeloid leukemia. Although cellular function of MN1 oncoprotein

Fig. 2. The gene expression maps of the chosen six genes for the SRBCT data. Each row corresponds to a gene, with the columns corresponding to different samples. Expression levels greater than the mean are shaded in red (green), and those below the mean are shaded in green (red). The genes ordered by hierarchical clustering. We used CLUSTER and TREEVIEW software which are publicly available at http://rana.lbl.gov

Fig. 3. The Box plots show the expression patterns of the chosen 6 genes for the SRBCT data, each numbered as IMAGE ID number.
T-cell leukemia/lymphoma 1 (TCL1) 2.66 O
Mal, T-cell differentiation
Human T-cell receptor

Transcription. This gene was also identified in Khan's work and to be involved in RAR/RXR-mediated transcription. BSS/WSS column present rank.

The gene list selected for the leukemia data

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Wald statistic</th>
<th>PAM</th>
<th>SVM-RFE</th>
<th>BSS/WSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystatin C (amyloid angiopathy and cerebral hemorrhage; CST3)</td>
<td>5.90</td>
<td>O</td>
<td>O</td>
<td>3</td>
</tr>
<tr>
<td>Mal, T-cell differentiation protein (MAL)</td>
<td>3.21</td>
<td>O</td>
<td>x</td>
<td>15</td>
</tr>
<tr>
<td>T-cell leukemia/lymphoma 1 (TCL1)</td>
<td>2.66</td>
<td>O</td>
<td>x</td>
<td>5</td>
</tr>
<tr>
<td>Human T-cell receptor active beta-chain (TCRB)</td>
<td>2.65</td>
<td>O</td>
<td>x</td>
<td>10</td>
</tr>
</tbody>
</table>

The abbreviations used is: the ratio of between classes sum of squares to within class sum of squares (BSS/WSS). ‘O’ denotes the genes selected by PAM and SVM-RFE methods. BSS/WSS column present rank.

3.2.2 Leukemia data

Leukemia dataset was composed of 7129 gene expression values in three classes of leukemias: B-cell and T-cell acute lymphoblastic leukemia (B-cell ALL-38 patients, T-cell ALL-9 patients) and acute myeloid leukemia (AML-25 patients) (Golub et al., 1999). As described in Dudoit et al. (2002), this dataset was preprocessed with thresholding, filtering and underwent a logarithmic transformation followed by standardization. The dataset consists of 38 training samples and 34 test samples.

SPM selected 4 genes (Table 5) from the 38 training samples. The training accuracy is 100% and test predictive accuracy is 91.17% (31/34), which means the SPM can predict the tumor classes for both seen and unseen samples using four genes with reasonably high accuracy. For comparison, we applied PAM and SVM-RFE methods to the leukemia data. PAM selected 10 genes from the 38 training samples and achieved an accuracy of 94.11% (32/34) and SVM-RFE selected 6 genes from the 38 training samples and achieved an accuracy of 91.17% (31/34). All six genes selected by SPM were also selected by PAM and one gene by SVM-RFE (Table 5).

Whether the selected genes serve as legitimate markers for cancer classification was further verified by applying a hierarchical clustering to the expression data of the selected genes. By visual inspection of the gene expression of the chosen four genes, three clusters were clearly separated (Fig. 5). Figure 6 illustrates three different patterns of the chosen four genes in the same fashion as Figure 2 in Lee and Lee (2003).

Previous studies already showed that cystatin C (CST3) gene was responsible for subtype classification of leukemia as a two-class (ALL/AML) problem (Golub et al., 1999). CST3 gene was also identified in Antonov et al.’s work for AML/ALL classification. But there was not any plausible conclusion about the biological function of this gene related to AML/ALL pathogenesis (Antonov et al., 2004). The relevance of MAL gene to T-cell ALLs was reported in the biological literature. The expression of MAL gene was shown to be significantly high in the PBMC of patients with T-cell ALL, as compared with that of chronic T-cell leukemia patients and normal subjects (Kohno et al., 2000). The MAL mRNA was expressed in T cells at intermediate and late stages of differentiation (Alonso et al., 1987), thyroid epithelial cells (Martin-Belmonte et al., 1998, Zachetti et al., 1995), and myelin-forming cells (Kim et al., 1995). MAL is a proteolipid that has been identified as a component of glycolipid-enriched membrane (Martin-Belmonte et al., 1998; Zachetti et al., 1995; Kim et al., 1995). MAL associates with glycosylphosphatidylinositol (GPI)-anchored proteins and the Src-like tyrosine kinase Lck in T lymphocytes. Cross-linking of GPlanchored proteins triggers signaling pathways leading to Lck activation and T-cell proliferation (Millan and Alonso, 1998; Shenoy-Scaria et al., 1992). Thus, the MAL molecule is closely associated with molecules for T-cell activation, and it probably has roles in activation and proliferation. The MAL gene was also identified in Antonov et al.’s work for AML/ALL classification (Antonov et al., 2004). We also identified some genes not identified in other works for AML/ALL classification. T-cell leukemia/lymphoma 1 (TCL1) gene and T-cell receptor active beta chain (TCRB) genes were up-regulated in B-type ALL and T-type ALL, respectively. TCL1 originally was cloned in 1994 as a gene on chromosome 14 involved in recurring chromosomal abnormalities of adult T-cell leukemia (Virgilio et al., 1993,1994). In these abnormalities, TCL1 was juxtaposed to a TCR locus and inappropriately regulated by the TCR cis-regulatory elements. TCL1 subsequently has been shown to be expressed in a variety of B-cell and T-cell neoplasm (Takizawa et al., 1998; Teitell et al., 1999; Narducci et al., 2000; Nakayama et al., 2000), but in neither hematopoietic progenitor cells (CD34+) nor mature lymphocytes. The increased expression of TCL1 was shown to be associated with the progression of immature B cell ALL (Fears et al., 2002). The rearrangement of the gene coding for the beta-chain of the T-cell receptor was found in T-cell leukemias and in T-cell lymphomas in biological literatures (O’Connor et al., 1985; Aisenberg et al., 1985; Bertness et al., 1985).

3.2.3 Other real data

We report the classification results and the size of gene sets for the three additional publicly available datasets: prostate cancer (Antonov et al., 2002), breast cancer (Wang et al., 2006) and ovarian cancer (Shen et al., 2002).

Table 5. The gene list selected for the leukemia data

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Wald statistic</th>
<th>PAM</th>
<th>SVM-RFE</th>
<th>BSS/WSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystatin C (amyloid angiopathy and cerebral hemorrhage; CST3)</td>
<td>5.90</td>
<td>O</td>
<td>O</td>
<td>3</td>
</tr>
<tr>
<td>Mal, T-cell differentiation protein (MAL)</td>
<td>3.21</td>
<td>O</td>
<td>x</td>
<td>15</td>
</tr>
<tr>
<td>T-cell leukemia/lymphoma 1 (TCL1)</td>
<td>2.66</td>
<td>O</td>
<td>x</td>
<td>5</td>
</tr>
<tr>
<td>Human T-cell receptor active beta-chain (TCRB)</td>
<td>2.65</td>
<td>O</td>
<td>x</td>
<td>10</td>
</tr>
</tbody>
</table>

The proportions of selecting each gene in SPM for 100 bootstrap samples. Red vertical lines denote the six genes selected for the SRBCT.
Fig. 5. The gene expression maps of the chosen four genes for the leukemia data. Each row corresponds to a gene, with the columns corresponding to different samples. Expression levels greater than the mean are shaded in red (green), and those below the mean are shaded in green (red). The genes ordered by hierarchical clustering. We used CLUSTER and TREEVIEW software which are publicly available at http://rana.lbl.gov.

Fig. 6. The Box plots show the expression patterns of the chosen four genes for the leukemia data, each numbered as Gene Accession number. (iii)X00437_s_at and (iv)X76223_s_at are identical expression patterns.

Table 6. Description of real datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample</th>
<th>Gene</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>62</td>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>62</td>
<td>4026</td>
<td>3</td>
</tr>
<tr>
<td>Brain</td>
<td>42</td>
<td>5597</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 7. The average number of genes selected from the real datasets

<table>
<thead>
<tr>
<th></th>
<th>Colon</th>
<th>Lymphoma</th>
<th>Brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPM</td>
<td>5.38</td>
<td>3.88</td>
<td>2.04</td>
</tr>
<tr>
<td>PAM</td>
<td>27.5</td>
<td>24.68</td>
<td>23.56</td>
</tr>
<tr>
<td>SVM-RFE</td>
<td>7.23</td>
<td>11.92</td>
<td>14.12</td>
</tr>
</tbody>
</table>
real datasets (Table 6) which are known to have relatively higher misclassification rates than the above two real datasets.

After preprocessing, all the data were base 10 log-transformed and the arrays were standardized. We performed 3-fold cross-validation and observed classification error rates. This procedure was repeated 50 times. We compared the test error rates and the average number of genes selected of our method with those of PAM and SVM-RFE.

Table 7 and Figure 7 display the average number of genes selected and boxplots of the misclassification rates of each classifier, respectively. For the Colon and Lymphoma data, SPM gave smaller misclassification rates with less average number of genes selected and for the Brain data, the misclassification rate of SPM was quite comparable with those of the other methods with less average number of selected genes. This result appears to show that our proposed method outperformed the other methods.

In order to show that variable selection may be helpful in improving prediction accuracy, we included SVM without gene selection for comparison. We used the R implementation svm() which is based on LIBSVM (Chang and Lin, 2001) and performed C-classification with Gaussian kernel. The parameter \( \sigma \) and the cost \( C \) were tuned by a grid search on \( \{2^{-12}, \ldots, 2^{12}\} \times \{2^{-5}, \ldots, 2^{10}\} \) by 10-fold cross-validation on each training dataset, similar to Myer et al. (2003) and Dettling (2004). As shown in Figure 7, SPM gave smaller misclassification rates than SVM without gene selection for these three datasets.

4. CONCLUSION AND DISCUSSION

The goal of this study is to provide an effective method for finding genes that accurately discriminate cancer subtypes. Our proposed methods using SPM gave a satisfactory classification performance with informative genes in the SRBCT and leukemia datasets. For the SRBCT data, SPM built an optimal class predictor consisted of six genes, that was able to assign each SRBCT sample to one of four subtypes, BL, EWS, NB and RMS, with 0% error rates. Our method required the smallest set of genes that can accurately classify samples. Although our method gives a little lower accuracy (31/34) than PAM (32/34) for leukemia classification, our method is able to select a smaller set of genes with ability to successfully classify among samples. To select a minimally required set of genes associated with optimum predictive performance is more cost-effective in cancer classification. We found four genes that are able to assign leukemia samples to one of three classes (AML, B-cell ALLs and T-cell ALLs), while PAM found 10 genes as an optimal set. All four genes selected by our method are also selected by PAM.

The efficiency of our method in finding a relatively small number of predictive genes will facilitate the search for new diagnostic markers. The method efficiently finds and ranks genes that can discriminate one subtype of tumor from another. For SRBCTs and Leukemias analyzed here, the predictive genes are attractive candidates for markers in RNA-based diagnostic test or immunohistochemical staining. In addition, our method may be used to...
select genes that are most significantly correlated with drug sensitivity or resistance, and to predict responses to chemotherapy according to gene expression profiles.

ACKNOWLEDGEMENTS

The authors wish to thank Trevor Hastie and Ji Zhu for their helpful comments and referees who informed them of related works and provided constructive comments that greatly improved this article. The research of J.Y.K. was supported by Korea Research Foundation (MOEHRD, Basic Research Promotion Fund) (KRF-2005-70-C00020). Funding to pay the Open Access publication charges for this article was provided by Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-70-C00020).

Conflict of Interest: none declared.

REFERENCES


