A comparative study of discriminating human heart failure etiology using gene expression profiles

Xiaohong Huang\textsuperscript{1}, Wei Pan\textsuperscript{1,*}, Suzanne Grindle\textsuperscript{2}, Xinqiang Han\textsuperscript{3}, Yingjie Chen\textsuperscript{3}, Soon J. Park\textsuperscript{3}, Leslie W. Miller\textsuperscript{3}, Jennifer Hall\textsuperscript{3}

November 12, 2004; Revised April 12, 2005

\textsuperscript{1} Division of Biostatistics, School of Public Health, University of Minnesota
\textsuperscript{2} Bioinformatics Core, Cancer Center, University of Minnesota
\textsuperscript{3} Cardiovascular Division, Department of Medicine, Medical School, University of Minnesota

\textsuperscript{*} Corresponding author. A460 Mayo Building (MMC 303), A460 Mayo Building (MMC 303), Minneapolis, MN 55455-0378, USA. Email: weip@biostat.umn.edu, phone: (612)626-2705, fax: (612)626-0660.
A comparative study of discriminating human heart failure etiology using gene expression profiles

Abstract

Human heart failure is a complex disease that manifests from multiple genetic and environmental factors. Although ischemic and non-ischemic heart disease present clinically with many similar decreases in ventricular function, emerging work suggests that they are distinct diseases with different responses to therapy. The ability to distinguish between ischemic and non-ischemic heart failure may be essential to guide appropriate therapy and determine prognosis for successful treatment. In this paper we consider discriminating the etiologies of heart failure using gene expression libraries from two separate institutions. We apply five new statistical methods, including partial least squares, penalized partial least squares, LASSO, nearest shrunken centroids and random forest, to two real datasets and compare their performance for multiclass classification. It is found that the five statistical methods perform similarly on each of the two datasets. In a simulation study, it is confirmed that the five methods tend to have close performance, though the random forest seems to have a slight edge.

Key words: Classification; Microarray data; Partial least squares; Penalized regression; Prediction.
Introduction

Human heart failure is a complex disease diagnosed in over 500,000 American people every year, causing more than 250,000 deaths annually. It may arise from coronary atherosclerosis, exposure to toxins, infection, inflammation, valvular disease leading to volume/pressure overload, or an underlying genetic or idiopathic event (Levy et al., 1996; Lloyd-Jones, 2001; Nicol et al., 2000). Emerging work suggests the heterogeneity of heart failure. For example, patients with ischemic heart failure have decreased survival compared to the non-ischemic heart failure (Felker et al., 2000; Dries et al., 2001) and respond differently to therapies (Kittleson et al., 2003; Felker et al., 2003; Doval et al., 1994; Singh et al., 1995). Although benefits can be achieved using ischemic heart failure therapy for idiopathic heart failure (and vice versa), cure rates will be markedly diminished, and warranted toxicities problems will be encountered. It may be critical to distinguish these characteristically similar but clinically somewhat distinct heart failures, to better optimize therapy. The ability to distinguish between different etiologies of heart failure may be essential to guide appropriate therapy and determine prognosis for successful treatment.

A new approach to discriminating etiologies of heart failure is gene expression profiling using DNA microarray technology, which has been shown to be promising in the diagnosis of human diseases or subdiseases, especially in cancer (Golub et al., 1999; Hedenfalk et al., 2001; Tibshirani et al., 2002). Recent genomic studies by three separate groups have demonstrated a distinct etiology dependent genomic pattern in the failing heart (Aronow et al., 2001; Tan et al., 2002; Hwang et al., 2002; Kittleson et al, 2004). These studies offer hope that the microarray gene expression analysis could potentially add to conventional laboratory approaches to diagnose different underlying etiologies of heart failure while simultaneously enhance prognostic criteria. It was hypothesized that heart failure arising from different underlying etiologies present with different gene expression patterns and that these differences could be used as a diagnostic tool. Here we test the hypothesis with two human heart failure datasets from different institutions.
Sample classification with gene expression data is statistically challenging due to the “small n, large p” problem (West 2002): the number of samples n is much smaller than the number of genes or predictors p. In our first dataset, we have n = 30 and p > 20000. Many new statistical methods have been developed or adapted to face the challenge. With more and more new methods emerging and existing methods being adapted, it becomes increasingly compelling for practitioners to compare and assess their performance, but there are few such comparative studies (Dudoit et al, 2001; Pan and Huang 2003; Wu et al, 2003). Huang and Pan (2003) compared several methods, including partial least squares (PLS) (Wold et al, 1984), nearest shrunken centroid (SC) (Tibshirani et al, 2002), and a penalized PLS (PPLS), for binary classification of gene expression data. They found that these methods are competitive. More recently, some authors (e.g. Segal et al, 2003; Wu et al, 2003) have shown the promising performance of least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and random forest (RF) (Breiman 2001). It is our main goal to evaluate and compare these methods using two human heart failure datasets. For this purpose, we also extend the PPLS, originally proposed for binary classification, to multiclass classification. We found that the above five statistical methods perform similarly. Furthermore, our analysis stresses the importance of utilizing multiple datasets for classification purposes.

**Methods**

**Binary classifiers: PLS, PPLS and LASSO**

We first briefly review the three binary classifiers, which was first designed for regression and can be directly applied to two-class classification, even when the number of covariates (i.e. genes here) is much larger than the sample size.

We code the response variable (i.e. class label) as $Y = 1$ for class 1 and $Y = -1$ for class 2. Suppose that $x_i$ is the expression level of gene $i$, $i = 1, \ldots, p$ with $p$ as the total number of the genes, and that we have $n$ samples in the training data. A challenge is that we have $n << p$. 


The main idea of partial least squares (PLS) (Wold et al) is to seek a few linear combinations of \( x = (x_1, ..., x_p)^T \), \( z_j = \alpha_j^T x \) for \( j = 1, ..., m \), then apply ordinary least squares (OLS) to regress \( Y \) on \( z_j \)'s to obtain

\[
\hat{Y}_{PLS} = \beta_0 + \sum_{j=1}^{m} \beta_j z_j
\]

with \( \beta \)'s as OLS estimates. The key of course is how to form linear components \( z_j \)'s. It turns out that

\[
\alpha_j = \arg \max_{\alpha} \text{Corr}^2(y, X\alpha) \text{Var}(X\alpha)
\]

with the constraints \( ||\alpha|| = 1 \), \( \alpha_l^T S \alpha = 0 \) for \( l = 1, ..., j - 1 \), where \( y \) is the vector of observed \( Y \)'s (in the training data), \( X \) is the design matrix (i.e. matrix of observed \( x \)'s), and \( S \) is the sample covariance matrix of \( x \)'s (Frank and Friedman 1992). In practice, the number of linear components \( m \) has to be chosen, typically by a form of cross-validation, such as leave-one-out cross-validation (LOOCV), to minimize misclassification errors.

PPLS is a penalized regression method in the framework of PLS (Huang and Pan 2003; Huang et al 2004). Suppose that we have built a PLS linear model, which can be rewritten as:

\[
\hat{Y}_{PLS} = \gamma_0 + \sum_{i=1}^{p} \alpha_i x_i.
\]

Then we shrink the PLS coefficients by soft-thresholding (Donoho and Johnstone, 1994; Donoho, 1995)

\[
a_i' = \text{sign}(a_i)(|a_i| - \lambda)_+ \quad \text{for} \quad i = 1, \ldots, p
\]

where \( \text{sign}(a) = 1 \) if \( a \leq 0 \) and \( \text{sign}(a) = -1 \) if \( a < 0 \), \( \lambda \) is a shrinkage parameter to be determined, and \( f_+ = \max(f, 0) \). It is common that the shrinkage leads to many \( a_i' = 0 \), effectively eliminating gene \( i \) from the model, thus gene selection is automatically accomplished. Next we construct a linear component \( z = \sum_{i=1}^{p} a_i' x_i \). Finally a PPLS model is built by regressing \( Y \) on \( z \) using OLS

\[
\hat{Y}_{PPLS} = \gamma_0' + \gamma_1' z,
\]

4
which can be re-expressed as $\hat{Y}_{PPLS} = \gamma_0 + \gamma_i \sum_{i=1}^{p} a_i^t x_i$. The parameters involved in building a PPLS model, such as the shrinkage parameter $\lambda$ and the number of PLS components, are estimated by LOOCV. The goal is to choose the largest shrinkage parameter and the smallest number of PLS components for which the LOOCV misclassification error estimate is minimized.

The LASSO estimates (Tibshirani 1996) in a linear model

$$\hat{Y}_{LSO} = \hat{\theta}_0 + \sum_{i=1}^{p} \hat{\theta}_i x_i$$

are obtained by

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,LSO})^2$$

subject to $\sum_{i=1}^{p} |\theta_i| \leq t$, where $Y_i$ is the observed response for sample $i$ and $\hat{Y}_{i,LSO}$ is its LASSO estimate, $i = 1, ..., n$, and $t$ can be chosen by LOOCV. The constraint can often force many $\hat{\theta}_i = 0$, leading to gene selection.

Note that the class label (1/-1) for the response $Y$ is binary, but in any of the above binary classifiers, the response $Y$ is treated as a continuous variable and the estimate $\hat{Y}$ could be any real number. To predict the class of a new sample, we use $\text{sign}(\hat{Y})$: if the estimated response $\hat{Y}$ is greater than or equal to 0, then we classify it into class 1; otherwise, class 2. In particular, this direct use of PLS for binary classification is different from other approaches (e.g. Nguyen and Rocke, 2002; Ghosh, 2002, 2003; Ding and Gentleman, 2004; Fort and Lambert-Lacroix, 2005); a distinct advantage of our approach is its simplicity, e.g., avoiding convergence problems when two classes are perfectly separable, which is common in microarray data with a small sample size and a large number of genes.

**Multiclass classifiers: nearest shrunken centroids and random forest**

Nearest shrunken centroids (SC) is built on a diagonalized linear discriminant analysis (DLDA) (e.g. McLachlan 1992; Hastie et al, 2001). Suppose that we have $K$ classes, $\bar{x}_{ik}$ is the mean expression level of gene $i$ in class $k$ of the training samples, $s_i^2$ is the pooled sample variance of gene $i$ of the training samples, and $\pi_k$ is the prior probability of a new sample
being in class $k$. The DLDA rule for a new sample $x^* = (x_1^*, ..., x_p^*)^T$ is

$$\delta_k(x^*) = \frac{\sum_{i=1}^{P} (x_i^* - \bar{x}_{ik})^2}{s_i^2} - 2 \log \pi_k.$$  

SC is motivated from the observation that many of the genes will not be predictive of the class membership and should be eliminated from the above DLDA rule. Formally, define

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i} \quad \text{with} \quad m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}},$$

where $n_k$ is the number of training samples in class $k$, and $\bar{x}_i$ is the overall mean expression level of gene $i$ in all the training samples. Note that by the definition, we have $\bar{x}_{ik} = \bar{x}_i - m_k s_i d_{ik}$. Let

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$$

for all $i$ and $k$, where $\Delta$ is the shrinkage parameter to be chosen by LOOCV. Then substituting $\bar{x}_{ik}$ in the DLDA rule by $\bar{x}'_{ik} = \bar{x}_i - m_k s_i d'_{ik}$, we obtain a SC rule

$$\delta'_k(x^*) = \frac{\sum_{i=1}^{P} (x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2 \log \pi_k.$$  

The new sample $x^*$ is assigned to class $k_0$ such that $k_0 = \arg \min_k \delta'_k(x^*)$.

Note that, if $d'_{ik} = 0$, then $\bar{x}'_{ik} = \bar{x}_i$ and thus gene $i$ plays no role in classifying for class $k$. Hence SC effectively accomplishes gene selection by shrinkage.

Random forest (RF) (Breiman 2001) is an ensemble of classification trees (Breiman et al 1984; Zhang and Singer 1999), which have been shown to be useful in tumor classification with microarray data (e.g. Zhang et al, 2001). It is designed to improve over a single classification tree. There are two random aspects that help generate multiple classification trees in RF. First, a bootstrap sample is repeatedly drawn from the original training data and then used to build a classification tree. Second, in building a classification tree, rather than using the best splitting variable (i.e. gene here) from all the available variables at each node, it chooses the best from a small random subset of all the variables. Each tree is grown to the maximum and no pruning is pursued. To predict the class for a new sample, the sample is applied to each tree and each tree votes by giving its prediction, then the majority vote is
taken as the final prediction for the sample.

**Extending a binary classifier to multiclass classification**

Here we describe how a multi-class ($K > 2$) classification problem can be handled by a binary classifier, such as PLS, PPLS and LASSO. It is achieved by formulating a multi-class classification problem as multiple two-class classification problems. We consider two most popular approaches: one is to compare each class against all the others, and the other is to compare all possible pairs of classes. Applications of these two approaches can be found, among others, in Friedman (1996), Hastie and Tibshirani (1998), Allwein et al. (2000), Ramaswamy et al (2001) and Dettling and Bühlmann (2003). In particular, Tan et al (2004) have considered the first approach for PLS.

The *one-against-others* approach is to reduce a $K$-class classification task to $K$ two-class classification problems. Formally, a new response is defined in the $k_{th}$ binary problem as:

$$Y^{(k)} = \begin{cases} 
1 & \text{if the sample is in class } k \\
-1 & \text{else}
\end{cases}$$

for $k = 1, \ldots, K$. Then we build $K$ binary classifiers. To predict a new sample with gene expression profile $x^*$, we apply $x^*$ to each binary classifier and yield $\hat{Y}_k$. Finally, the class of the new sample $C(x^*)$ is predicted as

$$C(x^*) = \arg\max_k \{\hat{Y}_k, \; k = 1, \ldots, K\}.$$  

That is, the new sample is classified into the class maximizing $\hat{Y}_k$.

The *pair-wise* approach reduces a $K$-class classification to $K(K-1)/2$ two-class classification problems (Friedman 1996). Specifically, for each of all possible pairs of classes, solve each of the two-class problems and then, for a new sample, combine all the pairwise decisions to form a K-class decision. Suppose that the new binary response in a pairwise comparison with classes $k_1$ and $k_2$ (with $1 \leq k_1 < k_2 \leq K$) is defined as

$$Y^{(k_1,k_2)} = \begin{cases} 
1 & \text{if the sample is in class } k_1 \\
-1 & \text{if the sample is in class } k_2.
\end{cases}$$
We build a binary classifier with response $Y^{(k_1,k_2)}$ using only samples belonging to class $k_1$ or $k_2$, and denote the fitted response value for a new sample with expression profile $x^*$ as $\hat{Y}_{k_1,k_2}$. As described earlier, we classify the new sample into class $k_1$ or $k_2$ according to the sign of $\hat{Y}_{k_1,k_2}$. After this is done for any $1 \leq k_1 < k_2 \leq K$, the final decision is to assign the new sample to the class that wins the most pairwise comparisons. In the case when there are multiple winning classes, we randomly pick one of the winning classes to be the final winning class. Comparing to the one-against-others approach, the pair-wise approach is computationally more expensive if $K \geq 4$.

**Gene ranking**

To explore the effect of the number of genes a model starts with on the classification performance, we have a preliminary gene ranking using a usual F-statistic. This univariate ranking is used throughout, and obviously is by no means to be optimal. For the purpose of the presentation in this section, we only need to consider a given gene. Suppose $x_{ik}$ is the gene expression level of the gene in sample $i$ that is in class $k$, $i = 1, \ldots, n_k$, and $k = 1, \ldots, K$, where $n_k$ is the total number of samples in class $k$ and $K$ is the total number of classes. Let $\bar{x}_k$ be the mean expression level of class-$k$ samples, $\bar{x}_.$ be the overall mean (across all the samples) and $N = \sum_{k=1}^{K} n_k$ be the total number of samples. We can construct an F-statistic as the ratio of the mean sums of squares for between-class and within-class variations:

$$F = \frac{MS_{\text{class}}}{MS_{\text{error}}} = \frac{\left(\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x}_.)^2\right)/(K-1)}{\left(\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2\right)/(N-K)}.$$  

We can rank all the genes based on their corresponding F-statistics: a gene with a larger F-statistic indicates a stronger relationship between its expression levels and the class membership in the samples, and therefore has a higher rank as a potential predictor of the class. We started with various models by including different numbers of top ranked genes. We considered models starting from the top 50, 100, 200, 400, 800, 1600, 3200, 6400, 9200, 12800, 16000, 19200, and all (22283 and 22277 for the two datasets respectively) genes respectively.
It is an incorrect practice in microarray experiments to first select genes using all the samples and then perform cross-validation using the selected genes, which gives downward biased prediction error estimates (Ambroise and McLachlan, 2002; Simon et al., 2003). Hence, it is essential to perform cross-validation on the entire model building process, including gene selection. In our study, we did honest cross-validation. In particular, we cross-validated gene selection (and other aspects of model building, such as parameter selection and estimation). Specifically, in LOOCV, we remove each sample from the data in turn (which is then treated as the test sample), carry out gene selection using F-statistic based on the remaining samples, build a classifier with the selected genes using the remaining samples, and then test the classifier on the left-out sample.

**Minnesota data**

Myocardial tissue samples from the left ventricular apex of patients with severe refractory heart failure were collected at the time of the left ventricular assist device (LVAD) placement at the University of Minnesota Medical School. The LVAD is a mechanical device that replaces the pumping of the failing ventricle. A total of 30 tissue samples were processed for microarray analysis on the Affymetrix Human Genome U133A chip containing \( \sim 22,000 \) genes. The initial data analysis was completed using Affymetrix Microarray Suite (MAS 5.0). A more complete description on the data is provided in Hall et al. (2004).

The heart patients are divided into three classes according to the underlying etiology. Patients with clinical ECHO and EKG evidence, history of previous myocardial infarctions, and direct observation of the heart for confirmation of infarction at the time of LVAD implantation are defined as ischemic. Patients with an ischemic etiology were further divided into two classes: patients with ischemia but without acute myocardial infarctions (ischemic class) and patients with ischemia that have had an acute myocardial infarctions within ten days of LVAD implant (IM class). the remaining patients were assigned to the idiopathic class. Among the total 30 samples, 10 of them are ischemic, 7 are IM and 13 are idiopathic.
PGA data
The PGA data were obtained in another heart failure study conducted at the Cardio-
Genomics PGA (Programs for Genomic Applications) at the Harvard Medical School. My-
ocardial samples were collected from patients undergoing heart transplantation whose fail-
ure arises from different etiologies (e.g. idiopathic, ischemic, alcoholic, valvular, and hyper-
trophic) and from normal organ donors whose hearts were not used for transplants. The tran-
scriptional profile of the mRNA in these samples was also measured with Affymetrix oligonu-
cleotide microarray technology. HG-U133 plus 2 chips containing 54,675 probe sets were used
and data were analyzed in MAS 5.0. In the PGA data set there were 11 normal samples, 11
ischemic samples and 14 idiopathic samples. The PGA dataset is publicly available at Ge-
nomics of Cardiovascular Development, Adaptation, and Remodelling. NHLBI Program for

In order to make the results comparable to those based on the HG-U133A chips used
in the Minnesota data, we matched the probe sets on a HG-U133 plus 2 chip with those
on a HG-U133A chip. Only six probe sets on the HG-U133 plus 2 chip could not be found
on a HG-U133A chip. Hence we used the remaining 22277 (= 22283 − 6) probe sets in the
following analyses with the PGA data.

Data preprocessing
To facilitate the application of penalized regression (i.e. PPLS and LASSO) so that their
regression coefficients are in the same unit and thus can be penalized using a global penalty
parameter, the expression levels of each gene were scaled to have sample variance 1.

Evaluations
In addition to PLS/PPLS, we will consider the shrunken centroids (SC) method (Tibshirani
et al., 2002), the LASSO (Tibshirani, 1996), and the random forest (RF) (Breiman 2001).
SC, LASSO and RF have been implemented in R (Ihaka and Gentleman, 1996), and are
easy to use; we applied their R functions using default parameter settings. SC and RF are
directly applicable to multiclass classification while LASSO, as PLS/PPLS, is itself a binary classifier. For multi-class classification with PLS and LASSO, we used the same approaches as described for PPLS.

Due to the small sample size (about 10 in each class) in each dataset, we use the leave-one-out cross-validation (LOOCV) to estimate the prediction error for each of the methods. Within this first-level LOOCV, a second-level LOOCV is used to select tuning parameters for each method to minimize cross validation errors. Specifically, in PLS, the smallest number of PLS components is selected among the PLS models that give the minimum LOOCV error. In PPLS, among the models with minimum LOOCV error, we first pick the ones with the smallest number of PLS components, then pick the one with the largest shrinkage parameter. In the SC method, the largest shrinkage is selected among the models that minimize the LOOCV error. The number of candidate threshold values and the number of cross validation folds are both set to be default (i.e. 30 and the smallest class size respectively). In LASSO, the maximum fraction parameter of the models that minimize LOOCV error is selected while the number of the candidate fraction values is set to be 51 (equally spaced from 0 to 1) and the number of cross validation folds is set to be the total sample size. In RF, every parameter is set to be default. For example, the number of trees to grow is set to 500, and the number of candidate splitting variables considered at each split is set to $\sqrt{p}$ by default, where $p$ is the total number of variables (i.e. genes).

Results

Classification with Minnesota data: One-against-others approach
Table 1 reports the LOOCV misclassification errors of the five classification methods for the Minnesota data with models starting from different numbers of top ranked genes ranging from 50 to all genes. Four kinds of errors are reported for each method: three one-against-others two-class LOOCV misclassification errors (Ischemic vs. others, IM vs. others and Idiopathic vs. others) and one three-class LOOCV misclassification error.
Table 1: LOOCV errors for three-class classification with Minnesota data: all patients.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>5ch vs other</th>
<th>IM vs other</th>
<th>1ch vs other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLS</td>
<td>PLS</td>
<td>SC</td>
<td>LSO</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>100</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>400</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>800</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>1200</td>
<td>10</td>
<td>12</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>1600</td>
<td>12</td>
<td>12</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>19200</td>
<td>13</td>
<td>12</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>22283</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>

Based on Table 1, we first note that the performances of any classifier is sensitive to the number of top ranked genes one starts with. For example, in the three-class classification, LASSO made 8 errors when only the top 200 genes are considered but made 18 errors when all 22283 genes are used. But there is no obvious relationship between the gene subset size and the five methods’ performances.

In the two-class classification problem of ischemic vs. others, PPLS, PLS, SC, LASSO and RF yielded LOOCV errors range from 7 to 15, 6 to 13, 8 to 13, 9 to 15 and 10 to 12 respectively. The five methods obtained similar numbers of errors in all instances. In the two-class classification problems of IM vs. others and idiopathic vs. others, all five methods yielded similar numbers of errors. In the overall three-class classification, again all the five classification methods perform similarly. There is no clear evidence that one method is clearly superior to others.

In order to assess whether a gene expression profile is affected by gender, we classified the 23 samples from male patients only. Among the 23 samples for male patients, 9 of them are ischemic, 7 are IM and 7 are idiopathic.

Table 2 reports the LOOCV misclassification errors of these five methods for the Min-
nesota data with males only. Again, the models start from different numbers of top ranked genes. The three two-class LOOCV misclassification errors and one three-class LOOCV misclassification error are estimated for each method as described before.

Based on Table 2, again we note that the classification performances of PPLS, PLS, SC, LASSO and RF can be quite sensitive to the number of top ranked genes one uses. For example, in the one-against-others two-class classification problem of ischemic vs. others, PPLS made 10 errors when the top 400 genes are considered but the number of errors suddenly increases to 14 when the top 800 genes are used. Again there is no obvious relationship between the gene subset size and the five methods’ performances.

Comparing PPLS, PLS, SC, LASSO and RF, we find that the five methods perform very similarly in almost all instances. There is no evidence that any method is clearly the best. One thing we noticed about LASSO is that when the model contains many genes, say top 12800, then it gives more errors than PPLS, PLS, SC and RF in the three-class classification. This could happen by chance since we did not observe the same trend in the decomposed one-against-others binary classifications.

As in Table 1, the results in Table 2 suggest that discriminating ischemic group from the other two groups was less accurate than distinguishing IM from the other two groups and separating idiopathic group from the other two groups.

If we compare Table 1 (30 patients) with Table 2 (23 male patients), we can see that the misclassification error rates of all the five methods in Table 2 are much higher than those in Table 1. Reduced sample size is likely a factor.

To see whether the high prediction errors are due to the presence of the three classes, we considered a binary classification problem. We applied all the five methods to the 10 ischemic and 13 idiopathic samples. We also assessed the classification accuracy on the male patients with 9 ischemic and 7 idiopathic samples. The classification results were shown in Table 3.

From Table 3, we can see that all the five methods have very similar performances in classifying ischemic and idiopathic samples. If we compare the classification performances of
Table 2: LOOCV errors for three-class classification with Minnesota data: males only.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>1st vs other</th>
<th>1M vs other</th>
<th>1st vs other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>PLS</td>
<td>SC</td>
<td>LSO</td>
<td>RF</td>
</tr>
<tr>
<td>50</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>200</td>
<td>12</td>
<td>12</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>400</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>800</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>1600</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>3200</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>6400</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>9600</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>12800</td>
<td>9</td>
<td>10</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>16000</td>
<td>13</td>
<td>12</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>19200</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>22483</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

These five methods with/without female patients, taking the sample size into consideration, we can see that the misclassification error rates with male patients only is much higher than those with all patients. This again is probably because the sample size (16) with males only is smaller. In particular, we noticed that LASSO is more sensitive to the small sample size.

Other models
In the previous classification problems, we only included linear terms of gene expression levels in a model. We also considered expanded models including squared terms of each gene’s expression levels. The motivation is to possibly improve model fitting, for instance, to avoid masking in linear models (Hastie et al, 2001). In this way, the number of variables in the new data is doubled (the original variables plus their squared terms) and we have 44566 variables. We repeated all the previous classification procedures and found that the classification performance did not improve (results not shown).

Classification with Minnesota data: pair-wise approach
We repeated the three-class classification with PPLS via the pair-wise approach. The results are included in Table 4. We assessed the classification of PPLS by including all 30
Table 3: LOOCV errors for two-class classification with Minnesota data: ischemic vs idiopathic.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>All (23 samples)</th>
<th>Males (16 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPLS</td>
<td>PLS</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>200</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>400</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>800</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>1600</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>3200</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>6400</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>9500</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>12800</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>16000</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>19200</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>22283</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

patients and with 23 male only. By comparing Table 4 to Table 1 - 2, we can see that the PPLS via one-against-others approach gives much smaller errors. This suggests that for this specific problem, the one-against-others approach is probably better. Again, we see the classification with male patients only gives much larger LOOCV misclassification error rates.

**Classification with PGA data: one-against-others approach**

Table 5 reports the LOOCV misclassification errors of the five methods for the PGA data. Four kinds of errors are reported for each method: three one-against-others two-class LOOCV misclassification errors (Normal vs. others, Ischemic vs. others, and Idiopathic vs. others) and one three-class LOOCV misclassification error.

Based on Table 5, we can see that the classification performances of PPLS, PLS, SC, LASSO and RF are quite stable with different numbers of top ranked genes one uses. In the two-class classification problem of normal vs. others, the five classification methods almost perform perfectly, where PPLS, PLS and RF have 0 errors in all circumstances, SC has 0
Table 4: LOOCV errors for three-class classification: PPLS with pair-wise approach.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>Minnesota data</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All patients</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>18</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>13</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>16</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
<td>17</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td></td>
<td>18</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td></td>
<td>17</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3200</td>
<td></td>
<td>16</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td></td>
<td>15</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9600</td>
<td></td>
<td>17</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12800</td>
<td></td>
<td>17</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td></td>
<td>19</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19200</td>
<td></td>
<td>14</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22283</td>
<td></td>
<td>19</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: LOOCV errors for three-class classification with PGA data: all patients.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>Normal vs other</th>
<th></th>
<th>Inh vs other</th>
<th></th>
<th>Ind vs other</th>
<th></th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPLS</td>
<td>PLS</td>
<td>SC</td>
<td>LSQ</td>
<td>RF</td>
<td>PPLS</td>
<td>PLS</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>400</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1600</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3200</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6400</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>9600</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>16000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>19200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>22283</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 6: LOOCV errors for two-class classification with PGA data: ischemic vs idiopathic.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>PPLS</th>
<th>PLS</th>
<th>SC</th>
<th>LASSO</th>
<th>RF</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3200</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9600</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12800</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>19200</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>22277</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

errors in all situations except for 3 cases and LASSO has 1 error in one case and 0 errors in all other cases. In the two-class classification problems of ischemic vs. others and idiopathic vs. others, PPLS, PLS, SC, LASSO and RF yielded 1-5 errors (mostly with 1-3 errors). That the problem of distinguishing normal from the others is the easiest confirms that the normal class is more separable from the other two classes. As for the three-class classification, the errors range from 1 to 2 and it is almost perfect.

It may be argued that a classification involving normal samples should be much easier because the normal class is very different from the other two classes. Correct diagnosis between ischemic and idiopathic would be much more challenging. Hence we conducted a two-class classification with 11 ischemic and 14 idiopathic samples. The results are shown in Table 6. The five methods perform almost perfectly: the misclassification errors range from 1 to 3 for all five methods in all models.

**Genes identified**

We consider genes remaining in a final model for each method. To save space, we restrict attention to models starting with all the genes for binary discrimination between ischemic
and idiopathic samples. Briefly, LOOCV was first used to select any tuning parameters in
a method (e.g. number of components in a PLS model), then a model with the selected
parameters was fitted using all the samples. Except that all the genes are used in a final
PLS model, for any of the other methods there may be fewer genes remaining in the final
model. In particular, LASSO can select at most $n$ genes with $n$ the number of the samples.
It turned out that random forest also used many of the genes.

Tables 7 and 8 lists the genes selected by at least four or three methods for the Minnesota
data and PGA data respectively. It can be seen that there is no overlap at all between the
two gene lists. Although the same genes are not identified from the two datasets, it is
clear that the beta-adrenergic signalling pathway may be involved given the inclusion of
CREM in the Minnesota data and AKAP6 in the PGA data. Furthermore, the inclusion
of metabolic-related genes, such as ATPase and GAPD, is not surprising given the class of
ischemic tissue.

We also give the univariate ranks of the genes (based on the F-statistics for the two
classes) in the above two tables. It shows that the two sets of genes (or more generally, the
genome in a final model) may not include some genes ranked high in the univariate ranking
while including some ranked low, highlighting a possible limitation of solely depending on
univariate ranks to select important genes.

More numerical results

PLS plots

To further explore why the methods all work much better for the PGA data than for
the Minnesota data, we drew some plots using the first and the second PLS components
for binary classification of ischemic vs idiopathic groups. We found that, for both datasets,
there was a clear separation between the two groups. However, in LOOCV, although again
the two groups were separable for both datasets, the left-out sample was more likely to be
closer to the other group than to its true group for the Minnesota data, leading to a high
LOOCV error rate; Figure 1 gives two examples. In contrast, in the PGA data, a left-out
sample tended to be close to its true group, resulting in a low LOOCV error rate. For more
Table 7: Genes selected by ≥ 4 methods in two-class classification (ischemic vs idiopathic) with Minnesota data. The last row gives the total numbers of the genes in the final models.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene</th>
<th>Rank</th>
<th>PPLS</th>
<th>PLS</th>
<th>SC</th>
<th>LSO</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>215066</td>
<td>PTPRP: protein tyrosine phosphatase, receptor type, F</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212005</td>
<td>UBX.D2: UBX domain containing</td>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>217234</td>
<td>VIL2: villin 2 (earin)</td>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>207092</td>
<td>BART1: binder of Arf Two</td>
<td>5</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>213831</td>
<td>NLK: nemo-like kinase</td>
<td>7</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212062</td>
<td>ATP9A: ATPase, Class II, type 9A</td>
<td>9</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>218008</td>
<td>FLJ22378: hypothetical protein FLJ22378</td>
<td>13</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212093</td>
<td>MTSG1: mitochondrial tumor suppressor gene 1</td>
<td>14</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>214543</td>
<td>QKL1: quaking homolog, KH domain RNA binding (mouse)</td>
<td>24</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>209487</td>
<td>RBPMS: RNA binding protein with multiple splicing</td>
<td>26</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>202877</td>
<td>CIQR1: complement component 1, q subcomponent, receptor 1</td>
<td>31</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>64438</td>
<td>FLJ22222: hypothetical protein FLJ22222</td>
<td>40</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>221928</td>
<td>LOC283445: hypothetical protein LOC283445</td>
<td>44</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212556</td>
<td>SCRIB: scrabble</td>
<td>48</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>202541</td>
<td>ARL3: ADP-ribosylation factor-like 3</td>
<td>50</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>200587</td>
<td>CLIC4: chloride intracellular channel 4</td>
<td>64</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>216231</td>
<td>Homo sapiens transcribed sequence with strong similarity to protein pdb:3HLA (H.sapiens) B Chain B, Human Class I Histocompatibility Antigen A2.1 (HLA-A2.1 Human Leucocyte Antigen)</td>
<td>73</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>220477</td>
<td>C20orf30: chromosome 20 open reading frame 30</td>
<td>76</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>208879</td>
<td>C20orf14: chromosome 20 open reading frame 14</td>
<td>88</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>207630</td>
<td>CREM: cAMP responsive element modulator</td>
<td>107</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212117</td>
<td>ARHQ: ras homolog gene family, member Q</td>
<td>116</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212904</td>
<td>KIAA1185: KIAA1185 protein</td>
<td>121</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>M33197</td>
<td>GAPD: glyceraldehyde-3-phosphate dehydrogenase</td>
<td>161</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>213507</td>
<td>KPNB1: karyopherin (importin) beta 1</td>
<td>163</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>207627</td>
<td>TFCP2: transcription factor CP2</td>
<td>208</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Total          - | 275  | 22283 | 36 | 22 | 883 |
Table 8: Genes selected by $\geq 3$ methods in two-class classification (ischemic vs idiopathic) with PGA data. The last row gives the total numbers of the genes in the final models.

<table>
<thead>
<tr>
<th>Probe Set</th>
<th>Gene</th>
<th>Rank</th>
<th>PPLS</th>
<th>PLS</th>
<th>SC</th>
<th>LSO</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>206375_at</td>
<td>HSPB3: heat shock 27kDa protein 3</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>202430_at</td>
<td>PLSCR1: phospholipid scramblase 1</td>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>209948_at</td>
<td>KCNMB1: potassium large conductance calcium-activated channel, subfamily M, beta member 1</td>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>APFX-TrpX-5</td>
<td></td>
<td>4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>212829_at</td>
<td>KIAA0592: KIAA0592 protein</td>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>221415_at</td>
<td>MYCBP: c-myc binding protein</td>
<td>7</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>219099_at</td>
<td>Cl2orf5: chromosome 12 open reading frame 5</td>
<td>8</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>219383_at</td>
<td>FLJ14213: hypothetical protein FLJ14213</td>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>208841_at</td>
<td>VDAC3: voltage-dependent anion channel 3</td>
<td>11</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>205359_at</td>
<td>AKAP6: A kinase (PRKA) anchor protein 6</td>
<td>12</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>202324_at</td>
<td>GOCAP1: golgi complex associated protein 1, 60kDa</td>
<td>14</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>208736_at</td>
<td>ARPC3: actin related protein 2/3 complex, subunit 3, 21kDa</td>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>215700_at</td>
<td>CPNE6: copine VI (neuronal)</td>
<td>16</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>207600_at</td>
<td>KCNC3: potassium voltage-gated channel, Shaw-related subfamily, member 3</td>
<td>17</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>217386_at</td>
<td>SPS2: selenophosphate synthetase 2</td>
<td>18</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>200961_at</td>
<td>MYOZ2: myozenin 2</td>
<td>19</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>211476_at</td>
<td>CBLB: Cas-Br-M (murine) ectropic retroviral transforming sequence b</td>
<td>21</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>209682_at</td>
<td>EIF4EBP2: eukaryotic translation initiation factor 4E binding protein 2</td>
<td>22</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>208799_at</td>
<td>FLJ10232: hypothetical protein FLJ10232</td>
<td>23</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>210500_at</td>
<td>NICE-4: NICE-4 protein</td>
<td>24</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>216721_at</td>
<td>LOC253512: hypothetical protein LOC253512</td>
<td>25</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>206417_at</td>
<td>CSH1: chorionic somatomamotropin hormone 1 (placental lactogen)</td>
<td>26</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>210961_at</td>
<td>WSB1: SOCS box-containing WD protein SWIP-1</td>
<td>27</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>206598_at</td>
<td>INS: insulin</td>
<td>28</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>210293_at</td>
<td>PTD004: hypothetical protein PTD004</td>
<td>29</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>207431_at</td>
<td>DEGS: degenerative spermatocyte homolog, lipid desaturase (Drosophila)</td>
<td>30</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>205207_at</td>
<td>IL6: interleukin 6 (interferon, beta 2)</td>
<td>31</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>221775_at</td>
<td>RPL22: ribosomal protein L22</td>
<td>32</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>200897_at</td>
<td>KIAA0992: palladin</td>
<td>33</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

|          | total                              |      | 32  | 22277 | 7 | 22 | 548 |

20
details see Supplemental Materials.

Figure 1: PLS plots for two cases in LOOCV for the Minnesota data comparing ischemic vs idiopathic. In both cases, the new sample labeled as “N” (i.e. left-out sample in LOOCV), belonging to class 1 and 2 respectively in the top and the bottom panels, is closer to the other group different from its true class, leading to misclassifications.

Permutation tests

Because of high misclassification error rates with the Minnesota data, it is of interest to investigate whether there is any signal at all in the data. This can be accomplished by a permutation test that compares misclassification errors resulting from using the original data with that from randomly permuted data; a P-value is defined as the proportion of permuted datasets with misclassification errors fewer than that of the original data. To generate a randomlypermuted data, we randomly permute the class labels of the original data. Because all the methods have similar performance, we consider the nearest shrunken centroid method with the Minnesota data. Tables 9 and 10 summarize the results of misclassification errors for 50 randomly permuted datasets for three- and two-class classifications respectively. It
Table 9: LOOCV three-class misclassification errors with the original Minnesota data and the percentiles of LOOCV errors with 50 permuted datasets by SC.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>Original data</th>
<th>Permutated data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOOCV errors</td>
<td>P-value</td>
</tr>
<tr>
<td>50</td>
<td>11</td>
<td>0%</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>0%</td>
</tr>
<tr>
<td>400</td>
<td>14</td>
<td>2%</td>
</tr>
<tr>
<td>1600</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>6400</td>
<td>13</td>
<td>2%</td>
</tr>
</tbody>
</table>

can be seen that the misclassification errors based on the original data are fewer than that based on the permuted data, leading to small P-values. This implies that, although there are relatively high misclassification error rates with the Minnesota data, the methods perform significantly better than a random guess.

Simulations

We did a simulation study to further evaluate the performance of the five classification methods. To mimic the real data, simulated data were generated from either a fitted PPLS or a fitted LASSO model to the Minnesota data comparing ischemic vs idiopathic, each containing top 50 genes in the initial model. Specifically, suppose that \( \hat{Y}_i \) is the fitted response value for sample \( i \) based on the original Minnesota data using PPLS or LASSO. Note that \( \hat{Y}_i \) is a real number without being dichotomized yet. Suppose that \( Y_i = 1 \) or \(-1\) is the class label of sample \( i \) in the original data, and \( \sigma^2 = \sum_{i=1}^{23} (Y_i - \hat{Y}_i)^2 / 23 \). To generate a simulated data, we independently draw \( Y_i^b \) from Normal distributions \( N(\hat{Y}_i, \sigma) \) for \( i = 1, \ldots, 23 \) and \( b = 1, \ldots, 50 \). Then we apply each method to a simulated dataset \( \{(sign(Y_i^b), X_i) : i = 1, \ldots, 23\} \), where \( X_i \) is the gene expression profile of sample \( i \) in the original Minnesota data, and obtain fitted values \( \hat{Y}_i^b \); the resulting misclassification error number for dataset \( b \) is \( \sum_{i=1}^{23} I[sign(\hat{Y}_i^b) \neq sign(\hat{Y}_i)] \).
Table 10: LOOCV two-class misclassification errors with the original Minnesota data and the percentiles of LOOCV errors with 50 permuted datasets by SC: ischemic vs idiopathic.

<table>
<thead>
<tr>
<th># of top genes</th>
<th>Original data</th>
<th>Permutated data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOOCV errors</td>
<td>P-value</td>
</tr>
<tr>
<td>50</td>
<td>5</td>
<td>0%</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0%</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>6%</td>
</tr>
<tr>
<td>1600</td>
<td>8</td>
<td>8%</td>
</tr>
<tr>
<td>6400</td>
<td>9</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table 11 summarizes the distributions of the misclassification errors of each method based on 50 simulated data with either PPLS or LASSO as the true model. It can be seen that in general all the methods perform similarly, though random forest seems to be most stable and has a slight edge, and the performance of LASSO and nearest shrunken centroid may deteriorate as the number of the genes included in a model is increased. We also did other simulations with the true models starting from various numbers of top genes and various noise levels, and observed similar phenomena: for details see Supplemental Materials.

**Discussion**

With more and more statistical methods being proposed for discriminant analysis for gene expression data, it has become increasing important to compare and evaluate their performances with real data, as it has been done in other contexts (e.g. Titterington et al, 1981). Comparing the five new methods with each other using the two real datasets, we did not find anyone uniformly better than the others. This may be disappointing to someone who wishes to find the best statistical method. However, in the current application, the similar performance of all the five methods on each of the two datasets provides reassurance on the interesting observation that it is not equally easy to distinguish the different etiologies.
Table 11: Percentiles of misclassification errors from 50 simulated datasets for two-class classification. The true model is either PPLS or LASSO fitted with top 400 genes to the Minnesota data to compare ischemic vs idiopathic.

<table>
<thead>
<tr>
<th>Methods</th>
<th># of top genes</th>
<th>True model: PPLS</th>
<th>True model: LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean 0% 25% 50% 75% 100%</td>
<td>Mean 0% 25% 50% 75% 100%</td>
<td></td>
</tr>
<tr>
<td>PPLS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.54 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.6 1 1.25 2 3 7</td>
<td>1 1 1 1 1 1</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>2.58 0 2 2 3 7</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>2.62 0 2 2 3.75 7</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>2.6 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>PLS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.58 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.6 0 2 2 3 7</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>2.56 0 2 2 3 7</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>2.52 0 2 2 3 7</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>2.54 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>LASSO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>2.84 0 2 2 3 10</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>3.48 0 2 2 4.75 10</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>3.48 0 2 2 4.75 10</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>SC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.76 1 2 2 3 7</td>
<td>1 1 1 1 1 1</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.74 1 2 2 3 7</td>
<td>1 1 1 1 1 1</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>2.96 0 2 3 3.75 7</td>
<td>0.28 0 0 0 1 1</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>3.58 0 2 3 5 8</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>4.08 1 3 4 5 10</td>
<td>1 1 1 1 1 1</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>6400</td>
<td>2.48 0 2 2 3 6</td>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>
of heart failure using expression profiles in the two datasets.

Both the one-against-others approach and the pair-wise approach have been widely used in extending a binary classifier to multi-class settings. Our result suggests that, at least for the two datasets used here, the one-against-others approach is better, which was found to be true with support vector machines but in general should also depend on which binary classifier is used (Rifkin et al, 2003). We also have observed that any of the five methods may be sensitive to the number of genes being included. This is particularly relevant because, although all the five methods (and many other methods) can handle any large number of genes, this does not dismiss the potential importance of a user’s preliminary ranking and screening of genes. Of course, all our observations here are based on the two datasets without consideration of statistical variability, further studies are needed to validate these points.

An interesting finding of this work is that it is difficult to discriminate the different etiologies of human heart failure using one gene expression dataset, and at the same time, it is quite easy for the other dataset. A possible explanation is the different types of the microarray chips used: Affymetrix HG-U133A chips were used in the Minnesota study while Affymetrix HG-U133 plus 2 chips were used in the PGA study. Because the HG-U133 plus 2 chips contain more genes (or ESTs), to minimize the effects of using different genes, we only used the genes present in the Minnesota data and still yielded much better performance for the PGA data. In fact, we used all the genes in the PGA data and obtained similar results for the PGA data. Although we can say that the performance difference in the two datasets is not caused by different genes contained on a chip, we do not know whether the more recent HG-U133 plus 2 chips provide more reliable measurements on gene expression. In addition, quality control criteria for the inclusion of a chip were nearly identical between the two datasets. We would suspect that the performance difference may be the result of different patient populations and different study protocols (e.g. no clearly pre-specified patient inclusion/exclusion criteria). As discussed in Simon (2004), a key to validating any prognostic and diagnostic biomarkers is the use of data that can reflect the full range of clinical variability. It can be also argued that the performance should be weighted on
the complexity of the disease. Challenges with the current clinical discrimination of ischemic versus non-ischemic heart failure is indeed why defining potential gene expression biomarkers may be a helpful additional approach in this characterization. A recognized limitation of utilizing heart tissue to identify biomarkers is the difficulty of collecting tissue. In summary, these studies stress the importance of collaborating efforts to share tissue/data to strengthen the search for applicable biomarkers.

Acknowledgements

XH and WP were supported by NIH grant HL65462. JH was supported by an AHA grant, the Lillehei Heart Institute and the Minnesota Medical Foundation. The authors are grateful to two reviewers for helpful and constructive comments; in particular, section “Other numerical results” was added based on the reviewers’ comments.

References


Available at http://www.stat.duke.edu/papers/working-papers-02.html.

