Gene expression

Regularized ROC method for disease classification and biomarker selection with microarray data

Shuangge Ma[1] and Jian Huang[2],*

[1]Department of Biostatistics, University of Washington, Washington, USA and [2]Department of Statistics and Actuarial Science, Program in Public Health Genetics, University of Iowa, Iowa City, IA, USA

ABSTRACT

Motivation: An important application of microarrays is to discover genomic biomarkers, among tens of thousands of genes assayed, for disease classification. Thus there is a need for developing statistical methods that can efficiently use such high-throughput genomic data, select biomarkers with discriminant power and construct classification rules. The ROC (receiver operator characteristic) technique has been widely used in disease classification with low-dimensional biomarkers because (1) it does not assume a parametric form of the class probability as required for example in the logistic regression method; (2) it accommodates case–control designs and (3) it allows treating false positives and false negatives differently. However, due to computational difficulties, the ROC-based classification has not been used with microarray data. Moreover, the standard ROC technique does not incorporate built-in biomarker selection.

Results: We propose a novel method for biomarker selection and classification using the ROC technique for microarray data. The proposed method uses a sigmoid approximation to the area under the ROC curve as the objective function for classification and the threshold gradient descent regularization method for estimation and biomarker selection. Tuning parameter selection based on the *V*-fold cross validation and predictive performance evaluation are also investigated. The proposed approach is demonstrated with a simulation study, the Colon data and the Estrogen data. The proposed approach yields parsimonious models with excellent classification performance.

Availability: R code is available upon request.

Contact: jian@stat.uiowa.edu

1 INTRODUCTION

Microarray experiments that monitor gene expression levels associated with different disease phenotypes have become commonplace in biomedical research. Each DNA sequence represented in microarrays can be considered a potential biomarker. It is of special interest to identify biomarkers that can be used to predict the phenotype of a new subject. Classification using genomic data is challenging due to high dimensionality of data and relatively small number of observations. Moreover, although the number of genes assayed is large, there may be only a small number of biomarkers that are associated with variations of phenotypes. Biomarker selection is needed along with model estimation.

Several dimension reduction techniques have been employed in classification problems with genomic data, see e.g., partial least squares (PLS, Nguyen and Rocke, 2002), principal component regression (Ma *et al*., 2005) and singular value decomposition in the Bayesian framework (West *et al*., 2001; Spang *et al*., 2001), among others. Using low-dimensional projections of the covariates as surrogates for the true covariates, one may obtain estimators with better prediction performance. An alternative to dimension reduction techniques is to use methods that are capable of simultaneous biomarker selection and model fitting. This may be accomplished by using regularization methods, e.g. the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), the least angle regression (LARS, Efron *et al*., 2004) and the gradient directed regularization method (Friedman and Popescu, 2004). Ghosh and Chinnaiyan (2005) proposed a hybrid method by first constructing a quantitative score using LASSO for the binary disease status and then using this score in the linear discriminant analysis.

A unique feature of disease classification is that it is important to assess both false positive and false negative errors, since these two types of errors usually have different consequences. A common practice is to use the receiver operating characteristic (ROC) curve approach (Pepe, 2003; Pepe *et al*., 2004), where the classification performance can be measured by the area under the ROC curve (AUC). Advantages of the ROC technique include (1) it does not assume a parametric form of the class probability as required in the logistic regression method; (2) it is adaptable to outcome-dependent samplings, e.g. the case-control design, which are widely used in medical studies and (3) it is relatively straightforward to assign different 'costs' to false positives and false negatives (Pepe, 2003; Pepe *et al*., 2005). Therefore, it is desirable to select biomarkers and to build classification rule based on direct optimization of the AUC.

Pepe *et al*. (2005) provide a detailed study using the AUC as objective function for combining biomarkers in a low-dimensional setting. But their study does not address the problems of (1) how to apply their method to situations where the number of predictors is large, as in microarray studies and (2) how to do biomarker selection, especially considering that it is biologically reasonable to assume that only a fraction of genes are related to phenotypes. Previous computational solutions, including brutal search and the tree-based approach (Abrevaya, 1999), are practically impossible with high-dimensional data.

In this article, we consider biomarker selection and classification using the ROC technique with microarray data. Because of the

---

*To whom correspondence should be addressed.

computational difficulty in using the empirical AUC directly, we propose using a sigmoid approximation to the empirical AUC as the objective function. With the sigmoid AUC criterion, we apply the threshold gradient descent regularization (TGDR) method (Friedman and Popescu, 2004) for simultaneous biomarker selection and model fitting.

2. STOCHASTIC MODEL AND ROC CURVE

2.1 Stochastic model

Consider a study with n subjects, where the outcome $Y \in \{0,1\}$ is a binary random variable, e.g. Y may denote presence or absence of cancer. Without loss of generality, we refer to $Y = 1$ as the diseased class and $Y = 0$ as the healthy class. For the i-th subject, expression values of d genes are measured: $\mathbb{X}_i = (X_{i,1}, \ldots, X_{i,d})$. The sample size n is at most in the hundreds, whereas the number of genes monitored d is in the thousands. We model the relationship between $\mathbb{X}$ and the phenotype Y with the semiparametric single index model $P(Y = 1|\mathbb{X}) = G(\beta'\mathbb{X})$, where G is an unknown increasing link function, $\beta = (\beta_{(1)}, \ldots, \beta_{(d)})$ is the d-vector of unknown regression parameter and β' denotes its transpose. No parametric form of the link function G is assumed. Thus the model assumption here is weaker than that in the logistic and probit regression models in which the parametric forms of G are assumed. Because of the monotonicity assumption on G, the classification rule can be constructed based on the linear risk scores $\beta'\mathbb{X}$. For example, we classify $Y = 1$ if $\beta'\mathbb{X} > c$, for a cutoff c chosen based on the estimated ROC, otherwise $Y = 0$.

2.2 ROC curve

To evaluate the performance of a classifier based on the linear risk scores $\beta'\mathbb{X}$, we employ the widely used measures of classification accuracy in medicine, namely the true and false positive rates (TPR and FPR). Also known as sensitivity and $1-$specificity, respectively, TPR and FPR are defined as

$$\text{TPR}(c) = P(\beta'\mathbb{X} \geqslant c \mid Y = 1) \text{ and } \text{FPR}(c) = P(\beta'\mathbb{X} \geqslant c \mid Y = 0),$$

for any cutoff c. The TPR and FPR can be summarized by the ROC curve, which is a two-dimensional plot of $\{(\text{FPR}(c), \text{TPR}(c)): -\infty < c < \infty\}$. The ROC curve demonstrates the balance between the TPR and FPR. Classification rules that have $(\text{FPR}(c), \text{TPR}(c))$ close to $(0,1)$ indicate satisfactory discriminators, while those with $(\text{FPR}(c), \text{TPR}(c))$ near the $45°$ line cannot discriminate between the diseased and the healthy classes.

The overall performance of a classifier can be measured by the AUC. For the n subjects, denote $\mathbb{D}$ and $\mathbb{H}$ as the index sets for diseased and healthy subjects with sizes n_D and n_H, respectively. Let $\mathbb{X}^\text{D}$ denote the biomarkers of a diseased subject and $\mathbb{X}^\text{H}$ the biomarkers of a healthy subject. For any β and the corresponding ROC generated with the linear risk scores $\beta'\mathbb{X}$, the empirical AUC is

$$\text{AUC}(\beta) = \frac{1}{n_\text{D} n_\text{H}} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} I(\beta'\mathbb{X}_i - \beta'\mathbb{X}_j > 0), \tag{1}$$

where I is the indicator function. It is interesting to note that the empirical AUC has the same form as the Mann–Whitney statistic for two-sample problems, which is equivalent to the Wilcoxon rank statistic. The ROC estimator is defined as the maximizer of $AUC(\beta)$ (Pepe *et al.*, 2005). This estimator is a special case of the maximum

rank correlation (MRC) estimator of Han (1987), which is only identifiable up to a scale constant. Without loss of generality, we assume $|\beta_{(1)}| = 1$, where $\beta_{(1)}$ denotes the first component of β, i.e. the first biomarker is the 'anchor biomarker.' We suggest a simple way of determining the anchor biomarker in Section 5.

3 THE SIGMOID MRC ESTIMATOR

Since the objective function in (1) is not continuous, finding its maximizer is difficult, especially with high-dimensional $\mathbb{X}$. One way to overcome this difficulty is to approximate the discontinuous objective function by a smooth function. We propose using the sigmoid function $s(x) = 1/[1 + \exp(-x)]$ as an approximation to the indicator function in (1).

For large $|x|$, $s(x)$ is an excellent approximation to $I(x) = I(x > 0)$. However, for x near 0, this approximation is poor and will introduce systematic bias in the estimation of β. An effective way to reduce the bias is to introduce a data-dependent small positive number σ_n and use $s_n(x) = s(x/\sigma_n)$ to approximate $I(x)$. A similar approach was proposed by Horowitz (1992) in the context of maximum score estimator for the binary response model. We refer to the resulting $\hat{\beta}$ using the sigmoid approximation as the sigmoid MRC (SMRC) estimator, i.e.

$$\hat{\beta} = \text{argmax} \left\{ R_n(\beta) = \frac{1}{n_\text{D} n_\text{H}} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\} \tag{2}$$

We set $|\hat{\beta}_{(1)}| = 1$ for identifiability. Since the scaled sigmoid function is an excellent approximation to the indicator function, the SMRC estimator should be close to the ROC estimator and $R_n(\beta)$ should be close to $AUC(\beta)$. The accuracy of the sigmoid approximation depends on σ_n. A rule of thumb for choosing σ_n is to guarantee all $|\beta'(\mathbb{X}_i - \mathbb{X}_j)/\sigma_n| > 5$ (Gammerman, 1996). A method for determining σ_n is given in Section 5. Empirical studies show that as long as σ_n is small enough, the estimates and the classification results are not sensitive to σ_n.

4 REGULARIZED ROC CLASSIFICATION

For simplicity of notation and without loss of generality, we assume that the first biomarker is the anchor and $\beta_{(1)} = 1$, and we still use β to denote the remaining coefficients $(\beta_{(2)}, \ldots, \beta_{(d)})'$.

4.1 The TGDR algorithm

The TGDR approach first establishes a parameter path in the high-dimensional coefficient space using the gradient descent method, and then identifies the best model along the parameter path with certain cross validation techniques (Friedman and Popescu, 2004). Let $\beta(\nu)$ denote the parameter path index by $\nu \in [0, \infty)$. Let $\Delta\nu$ be the infinitesimal positive increment as in ordinary gradient descent methods (Friedman and Popescu, 2004). In the implementation of this algorithm, we choose $\Delta\nu = 1 \times 10^{-4}$. For any threshold $0 \leq \tau \leq 1$, the TGDR algorithm consists of the following iterative steps:

(1) Initialize $\beta(0) = 0$ and $\nu = 0$.

(2) Compute the negative gradient $g(\nu) = -\partial R_n(\beta)/\partial\beta$ evaluated at $\beta(\nu)$. Denote the j-th component of $g(\nu)$ as $g_j(\nu)$. If $\max_j\{|g_j(\nu)|\} = 0$, stop the iterations.

(3) Compute the vector $f(\nu)$ of length d, where the j-th component of $f(\nu)$: $f_j(\nu) = I\{|g_j(\nu)| \geq \tau \cdot \max_l|g_l(\nu)|\}$.

(4) Update $\beta(\nu + \Delta\nu) = \beta(\nu) + \Delta\nu \times g(\nu) \times f(\nu)$ and replace ν by $\nu + \Delta\nu$, where the product of f and g is component-wise.

(5) Steps 2–4 are repeated k times. The number of iterations k is determined by cross validation as described below.

When $\max_l\{|g_l(\nu)|\}$ is less than a prespecified criterion, the iteration can be stopped. We recommend tracking the magnitude of $\max_l\{|g_l(\nu)|\}$ and the plot of the cross validation function as a function of k to determine the number of iterations in Step 5.

Detailed discussions of the TGDR algorithm can be found in Friedman and Popescu (2004), where a graphic presentation is also available (Figures 1 and 3 therein). Stable estimates are expected with non-zero τ, since covariates with small gradients are excluded from the model. The tuning parameters τ and k jointly determine the property of $\hat{\beta}$. When $\tau \approx 0$, $\hat{\beta}$ is dense for all values of k. When $\tau \approx 1$, $\hat{\beta}$ is sparse for small k and remains so for a relatively large number of iterations, but will become dense eventually. At the extreme when $\tau = 1$, the TGDR usually increases in the direction of a single covariate in each iteration. This mimics the incremental forward stge-wise strategy described in Hastie *et al.* (2001). When τ is in the middle range, the characteristics of $\hat{\beta}$ are between those for $\tau = 0$ and $\tau = 1$.

In a linear regression model, Friedman and Popescu (2004) show that the TGDR can provide a path connecting the solutions roughly corresponding to the PLS/RR (ridge regression) and the solutions roughly corresponding to the LASSO by varying the thresholds. Moderate-to-large threshold values create paths that involve more diverse absolute coefficient values than the PLS/RR but less than the LASSO. Our numerical studies suggest that the conclusions drawn from linear regression are applicable here.

4.2 Tuning parameter selection

We propose using the following *V*-fold cross validation (Wahba 1990) to determine the tuning parameter k for a given τ. For a pre-defined integer V, partition the data randomly into V non-overlapping subsets of equal sizes. Choose k to maximize the cross-validated objective function

$$\text{CV score} = \sum_{v=1}^{V}[R_n(\hat{\boldsymbol{\beta}}^{(-v)}) - R_n^{(-v)}(\hat{\boldsymbol{\beta}}^{(-v)})], \qquad (3)$$

where $\hat{\boldsymbol{\beta}}^{(-v)}$ is the TGDR estimate of β based on the data without the *v*-th subset for a fixed k and $R_n^{(-v)}$ is the function R_n defined in (2) evaluated without the *v*-th subset. Since the CV score in (3) contains differences of biomarkers from the two phenotypes, the usual leave-out-one cross validation is not applicable here.

Since the performance of the TGDR estimators for different threshold values is of interest, we employ cross validation with respect to k only. We discuss the effect of different τ with empirical studies in Section 5. Related discussions can also be found in Gui and Li (2005). Beyond selecting the model (corresponding to the cross-validated tuning parameters) with the best predictive performance, the *V*-fold cross validation also provides partial protection against overfitting (Nguyen and Rocke, 2002).

4.3 Assessing prediction significance

4.3.1 *Observed predictive distribution of AUC*   A simple measure of classification performance is the empirical AUC itself. However, if the estimation and evaluation are based on the same

data, then the classification performance will in general be over estimated. Detailed discussion on related problems can be found in Ambroise and McLachlan (2002). Ideally, we would want to use independent data to assess prediction performance, but such data are usually not available. So we propose the following prediction performance evaluation based on random partition.

(1) We first partition the data randomly into a training set of size n_1 and a testing set of size n_2 with $n_1 + n_2 = n$. Dudoit *et al.* (2002) suggest $n_1 \sim 2/3n$.

(2) We use the training set to compute the SMRC estimator and to construct the classification rule based on this estimator. We then use this classification rule to predict the disease status for individuals in the testing set based on their gene expression profiles. The AUC is then computed for the testing set.

(3) To take into account the fact that we may get a large value of the AUC by chance with a 'lucky' partition, we repeat this process many (e.g. 1000) times. Each time a new partition is made and the value of the testing set AUC is computed.

With this procedure, we obtain a Monte Carlo estimation of the distribution of prediction AUC by partitioning the observed data. We call it the observed predictive distribution (OPD) of AUC, which provides an honest measure of the classification performance of the proposed methodology.

4.3.2 *Permutation predictive distribution*   We propose the following approach as a benchmark for assessing the significance of the prediction AUC, which is motivated by the idea of standard permutation method for hypothesis testing. In a permutation test, the value of the test statistic based on the observed data is first calculated. The significance of this observed test statistic is evaluated using the permutation distribution, which is the distribution of the test statistic calculated based on randomly permuted data. In the present problem, the OPD of AUC plays the role similar to the observed value of the test statistic in a standard hypothesis testing setting. We compare this OPD with the distribution of the AUC calculated using the uninformative data obtained by random permutation and partition as follows.

We first randomly permute the binary outcome Y, but keep the indices of the biomarkers fixed. We then couple the permuted outcome with the biomarkers. Specifically, let $(\pi_1, \ldots, \pi_n)$ be a permutation of $(1, \ldots, n)$. The permuted dataset is $(Y_{\pi i}, \mathbb{X}_i)$, $i = 1, \ldots, n$. We permute the data 1000 times. Each time, we partition the permuted data into a training set of size n_1 and a testing set of size n_2, and compute the testing set AUC in the same manner as described above. This yields the Monte Carlo distribution of the AUC with permuted data, i.e. the permutation predictive distribution (PPD) of AUC.

We note that calculations of OPD and PPD are parallel. The OPD is calculated from the partitioned observed data, whereas the PPD is calculated from the partitioned permuted data. Well-separated OPD and PPD distributions indicate that the model estimated with the proposed approach is effective in terms of prediction, whereas substantially overlapped distributions suggest that either the proposed approach is not effective or the biomarkers do not have good discriminant power. In addition to the AUCs, the classification error rates with the observed data and permuted uninformative data can also be computed and compared following the above procedure.

Table 1. Simulation study: means of AUC and classification error (with their standard errors in parentheses)

(n_H, n_D)	$\pi = 0.05$ Small change	Large change	$\pi = 0.5$ Small change	Large change
(15,15)				
AUC	0.90 (0.01)	0.89 (0.01)	0.90 (0.01)	0.90 (0.02)
Error	0.07 (0.10)	0.06 (0.07)	0.04 (0.09)	0.05 (0.07)
(20,10)				
AUC	0.92 (0.01)	0.93 (0.01)	0.93 (0.01)	0.92 (0.01)
Error	0.10 (0.12)	0.09 (0.11)	0.06 (0.09)	0.07 (0.10)
(50,50)				
AUC	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)
Error	0.09 (0.06)	0.07 (0.04)	0.06 (0.05)	0.05 (0.04)
(70,30)				
AUC	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)
Error	0.07 (0.05)	0.07 (0.05)	0.05 (0.04)	0.05 (0.06)

We note that the above approach is for evaluating prediction significance for a given method. For comparing different methods, we only need to compare their respective OPDs.

5 NUMERICAL STUDIES

5.1 Simulation study

We use the same simulation settings as in Ghosh and Chinnaiyan (2005, Table 1). We first generate $d = 1000$ dimensional vectors from two populations. We consider the following sample size combinations $(n_D, n_H) = (15, 15), (10, 20), (50, 50)$ and $(30, 70)$, where n_D and n_H denote the number of samples in the groups with $Y = 1$ and $Y = 0$, respectively. All the potential biomarkers are assumed to be independently normally distributed with variance 1. We assume a model in which a fraction π of the potential biomarkers are differentially expressed between the two classes. $\pi = 0.05$ and 0.5 are considered. Moreover, we investigate two scenarios. In the first scenario, there is a big change in expression in the differentially expressed biomarkers, a shift of 5 units in the mean. In the second scenario, the fold change is only 1.5 unit difference in mean. For each simulated dataset, the SMRC estimate is obtained from a randomly sampled training set of size $2/3 \times (n_D + n_H)$ and the AUC is computed with the testing set. Summary statistics based on 200 simulated datasets are given in Table 1.

It can be seen that in all simulated settings, the SMRC estimates are satisfactory in the sense that they have large AUCs and small classification errors. The AUCs increase as sample size increases, as expected. The patterns of this increase are quite consistent for different values of π and different values of change in gene expression between two groups.

5.2 Colon data and Estrogen data

Colon data. In this dataset, expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes are measured using the Affymetrix gene chip. A selection of 2000 genes with highest minimal intensity across the samples has been made by Alon *et al.* (1999), and these data are publicly available at http://microarray. princeton.edu/oncology/. The Colon data have been analyzed in several previous studies using other statistical approaches, see

e.g. Dettling and Buhlmann (2003), Pochet *et al.* (2004), Ben-Dor *et al.* (2000) and Nguyen and Rocke (2002).

Estrogen data. This dataset was first presented by West *et al.* (2001) and Spang *et al.* (2001). It includes expression values of 7129 genes of 49 breast tumor samples. The expression data were obtained using the Affymetrix gene chip technology and are available at http://mgm.duke.edu/genome/dna_micro/work/. The response describes the lymph nodal (LN) status, which is an indicator for the metastatic spread of the tumor, an important risk factor for disease outcome. Of the total, 25 samples are positive (LN+) and 24 samples are negative (LN−). We threshold the raw data with a floor of 100 and a ceiling of 16 000. Genes with max(expression)/min(expression) < 10 and/or max(expression) − min(expression) < 1000 are also excluded (Dudoit *et al.*, 2002). A base 2 logarithmic transformation is then applied. The Estrogen data have also been studied in Dettling and Buhlmann (2003).

We identify the anchor biomarker as follows. Compute the sample standard errors of the d biomarkers $se_{(1)}, \ldots, se_{(d)}$ and denote their median as med.se. Compute the adjusted standard errors as $0.5(se_{(1)} + \text{med.se}), \ldots, 0.5(se_{(d)} + \text{med.se})$. Then the biomarkers are ranked based on the t-statistics computed with the adjusted standard errors. This adjusted t-statistic is similar to a simple shrinkage method discussed in Cui *et al.* (2005). The biomarker with the largest absolute values of the adjusted t-statistic is chosen as the anchor biomarker. For the anchor biomarker, if the sample mean of the diseased class is larger, $\beta_{(1)} = 1$, otherwise $\beta_{(1)} = -1$.

After the anchor biomarker is chosen, we choose σ_n to be the largest number that satisfies $|(x_1^D - x_1^H)/\sigma_n| > 5$, where x_1^D and x_1^H denote the expression levels of the anchor biomarker for the diseased and healthy subjects. The rationale for this is as follows. If we only have the anchor biomarker, then this approach guarantees that all the terms x/σ_n have absolute values >5. Now consider multi-biomarker cases. If we neglect the correlation structures among biomarkers, then for biomarker j, $P(\beta_{(j)}(x_j^D - x_j^H)/\sigma_n > 0) > 0.5$. So when we add more biomarkers, the absolute values of $\beta'(\mathbb{X}^D - \mathbb{X}^H)/\sigma_n$ tend to become bigger. When we take the correlation structures into consideration, the above argument is not necessarily true for all biomarkers, but will remain true on average. For σ_n determined in this manner, we try out tuning parameters $\sigma_n \times 0.1, \sigma_n \times 0.01 \cdots$ to the same data. We find that the classification results are very stable.

The 500 genes with the largest absolute values of the adjusted t-statistics are used for classification. The remaining genes have relatively small change across the subjects. Note there is no computational limitation on how many genes can be used in the SMRC classification. We use 500 genes only to gain further stability. The genes are then standardized to have zero means and unit variances.

Five-fold cross validation is used for tuning parameter selection. We show in Table 2 the cross-validated k, number of genes with non-zero coefficients and the CV scores for each fixed τ. It can be seen that generally cross-validated k increases and the number of non-zero coefficients decreases as τ increases. However, the change of CV score is very small. Our extensive empirical studies show that in general more iterations are needed for larger τ, which will also lead to more parsimonious models. Parsimonious models are preferred when the CV scores are comparable. So we choose $\tau = 1.0$ for both datasets.

For comparison purposes, we also consider the logistic model for the two datasets, i.e. assuming G is the logit function. Considering

Table 2. Model features for different τ

τ	k	Colon variable	CV	k	Estrogen variable	CV
0.0	448	500	174.1	117	500	247.5
0.1	444	500	173.7	69	500	248.0
0.2	440	500	175.1	106	500	247.9
0.3	466	500	180.8	77	500	246.7
0.4	479	467	173.4	61	485	245.6
0.5	554	351	175.0	70	444	234.1
0.6	638	266	171.2	92	306	240.0
0.7	950	150	166.3	153	187	236.3
0.8	1410	74	172.0	284	104	236.1
0.9	2320	42	169.5	1280	55	231.8
1.0	4280	31	170.1	2630	24	230.0

k, cross-validated number of iterations; variable, number of genes with non-zero coefficients.

Table 3. Colon data: genes with non-zero coefficients

GeneID	SMRC	Logistic	GeneID	SMRC	Logistic
Hsa.750	0.57	–	Hsa.467	−0.58	–
Hsa.668	−0.46	–	Hsa.8147	−0.65	−0.22
Hsa.81	0.33	–	Hsa.43279	−0.71	–
Hsa.36689	–	−0.08	Hsa.37937	–	−0.05
Hsa.949	0.56	–	Hsa.8219	−0.72	–
Hsa.2487	0.96	–	Hsa.3306	–	0.26
Hsa.10047	0.67	–	Hsa.34914	0.60	–
Hsa.10706	−0.64	–	Hsa.2051	0.50	–
Hsa.8214	0.77	–	Hsa.3016	1.08	–
Hsa.5392	0.61	–	Hsa.1410	0.46	0.01
Hsa.1786	−0.79	–	Hsa.2808	−0.73	–
Hsa.24582	0.44	–	Hsa.43405	−1.04	–
Hsa.2928	0.54	0.07	Hsa.17426	1.09	–
Hsa.1454	−1.81	–	Hsa.627	1.00	0.08
Hsa.2688	0.28	–	Hsa.3254	−0.56	–
Hsa.6814	1.56	0.32	Hsa.2291	−0.83	–
Hsa.43331	−0.86	–	Hsa.6782	0.50	–

Table 4. Estrogen data: genes with non-zero coefficients

GeneID	SMRC	Logistic	GeneID	SMRC	Logistic
D37931_at	0.18	–	D86957_at	0.08	–
D87468_at	−0.36	–	HG2247_at	−0.39	−0.04
HG3431_s_at	0.26	–	HG4716_at	−0.36	−0.09
J02871_s_at	0.12	–	J03827_at	–	−0.09
L21998_at	−0.36	–	L26336_at	0.19	0.31
L40401_at	0.25	0.07	M14745_at	0.28	–
M26311_s_at	−0.27	–	M29874_s_at	0.04	–
M32053_at	0.43	–	M62403_s_at	–	0.01
M74093_at	−0.21	–	U01038_at	−0.33	−0.03
U42408_at	–	−0.12	U45955_at	−0.29	–
U84011_s_at	0.36	–	X03635_at	1.00	0.79
X13334_at	–	−0.12	X56667_at	−0.44	−0.08
X57809_at	−0.25	–	X84373_at	–	0.037
X86693_at	0.31	–	X87237_at	−0.45	−0.51
Z00010_at	−0.29	–	Z29083_at	–	0.10

that the sample sizes are much smaller than the number of covariates, we apply the LASSO regularization to the logistic regression. Tuning parameter in the LASSO is chosen using the same 5-fold cross validation method. The OPD and PPD AUCs for the logistic-LASSO method are computed in the same manner as described in Section 4.3.

In Tables 3 and 4, we list the genes with non-zero coefficients and the corresponding estimates for the SMRC approach with $\tau = 1.0$ and the logistic-LASSO method. Since the gene expressions have been normalized to have unit variances, the estimated coefficients are directly comparable. Larger absolute values of coefficients indicate stronger influences. The SMRC identifies more genes than the logistic-LASSO approach. A majority of genes identified by the logistic-LASSO model are also identified by the SMRC. When there are overlaps, the coefficients from two approaches have the same signs, which suggests similar biological conclusions.

Because of page limitation, we did not include the descriptions of all the genes in Tables 3 and 4, which can be found from the NCBI website (www.ncbi.nlm.nih.gov). For example, for the Colon data, the gene corresponding to Hsa.467 is MYL6. This gene is expressed in smooth muscle tissues. Muscle-specific genes were also found to be correlated with colon tumor by Alon *et al.* (1999). Hsa.81 (gene name: KLF5) is a positive regulator of cellular proliferation. Hsa.8147 (gene name: AKT1) is a critical mediator of growth factor-induced neuronal survival. Survival factors can suppress apoptosis in a transcription-independent manner by activating the serine/threonine kinase AKT1. Hsa.43279 (gene Name: TSPAN1) encodes a protein that mediates signal transduction events that play a role in the regulation of cell development, activation, growth and mobility. Hsa.8219 (gene name: CDKN1A) encodes a potent cyclin-dependent kinase inhibitor. The encoded protein binds to and inhibits the activity of cyclin-CDK2 or -CDK4 complexes, and thus functions as a regulator of cell cycle progression. The expression of this gene is tightly controlled by the tumor suppressor protein p53, through which this protein mediates the p53-dependent cell cycle G1 phase. Hsa.3016 (gene name: S100P) encodes a protein which is a member of the S100 protein family. S100 proteins are involved in the regulation of a number of cellular protein which may also play a role in the etiology of prostate cancer.

For the Estrogen data, D37931 (gene name: RNS4) encodes a component of the interferon-regulated 2-5A system that functions in the antiviral and antiproliferative roles of interferons. Mutations in this gene have been associated with predisposition to prostate cancer. X03635 (gene name: ESR1) is the estrogen receptor which is a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding and activation of transcription. M14745 (gene name: BBC2) encodes a protein that is a member of the BCL-2 family. BCL-2 family members are known to be regulators of programmed cell death. This protein positively regulates cell apoptosis by forming heterodimers with BCL-xL and BCL-2, and reversing their death repressor activity.

We note that although the genes identified for the two datasets exhibit strong predictive power of disease status as demonstrated above, the analysis here does not provide information on whether they are just genomic markers correlated with the disease status or are actually in the pathways leading to the diseases. Indeed, as is
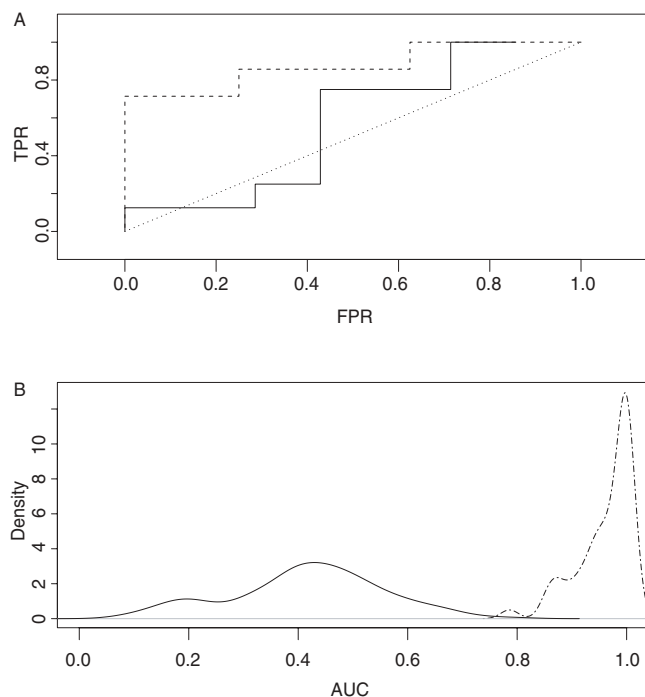
Fig. 1. Estrogen data. (**A**) representative plots of predictive ROCs. The dashed line is the predictive ROC based on the observed data. The solid line is the predictive ROC based on uninformative permuted data. The dotted line is the diagonal line. (**B**) kernel density estimations for the OPD (dashed line) and PPD (solid line) of AUCs.

typical with microarray results, further investigation based on independent assay and/or independent sample is desirable.

The classification performance of the proposed method for the two datasets is evaluated using the approach described in Section 4.3. Both the number of random partitions for the observed data and the number of permutations are 1000. For the Colon data, the SMRC has mean AUCs 0.91 (0.06) and 0.46 (0.12) for the OPD and PPD, respectively, where the values in the parentheses are the standard errors. Corresponding mean classification errors are 0.14 (0.07) for observed data and 0.44 (0.11) for randomly permuted data, respectively. For the Estrogen data, the SMRC has mean AUCs 0.96 (0.05) and 0.42 (0.14) for the OPD and PPD, respectively, with corresponding mean classification errors 0.06 (0.07) and 0.49 (0.12). In Figure 1A, we show representative plots of the ROC. The dashed line is the ROC based on observed data with AUC = 0.97, whereas the solid line is the ROC obtained based on the permuted uninformative data with AUC = 0.43, both based on one run of partition only. In Figure 1B, we show the density plots of the OPD and PPD of AUC with 1000 partitions for the Estrogen data. We see that the two distributions are well separated. A Wilcoxon test of the difference of these two distributions gives the P-value < 0.0001. Corresponding plots for the Colon data are similar and omitted.

For the Colon data, the logistic-LASSO method yields mean AUC 0.88 (0.08) and mean classification error 0.17 (0.08). For the same dataset and using the 200 top ranked genes based on the marginal Wilcoxon rank test, Dettling and Buhlmann (2003, Table 1) shows the mean classification errors are 0.16 (LogitBoost, 100 iterations), 0.18 (AdaBoost), 0.18 (1-nearest-neighbor) and

0.15 (classification tree). In Pochet *et al.* (2004, Supplementary information), 2000 genes are used and the SVM-based methods have mean AUCs 0.85 and mean classification errors 0.18, whereas the principal component analysis-based approaches have even smaller AUCs and larger classification errors. So for the Colon data, the proposed SMRC performs the best in terms of AUC and classification error. We also note that since different sets of genes are used in Dettling and Buhlmann (2003) and Pochet *et al.* (2004), the above results only provide a rough comparison.

For the Estrogen data, the logistic-LASSO model has mean AUC 0.92 (0.07) and mean classification error 0.12 (0.08) for the observed data, which suggests less satisfactory classification performance. Dettling and Buhlmann (2003, Table 1) uses 200 genes selected based on the marginal Wilcoxon tests and yields classification errors 0.04, 0.06 and 0.08 using different LogitBoost methods, classification error 0.04 for AdaBoost, 0.14 for 1-nearest-neighbor approach and 0.04 by using classification tree.

6 CONCLUDING REMARKS

It is of practical importance to develop computationally feasible models and methods for biomarker selection and disease classification with high-dimensional genomic data. In this article, we adapt the ROC technique, which has been successfully used with low-dimensional data, to genomic studies. We propose using the smooth sigmoid objective function as an approximation to the discontinuous AUC, so that it is computationally feasible for high-dimensional genomic data. The TGDR method, which is firstly developed for linear regression models, is adapted for biomarker selection and regularization. Applications of the proposed method to a simulation study and two studies using Affymetrix gene chip data suggest that it can select biomarkers with satisfactory classification performance as measured by AUC via cross validation.

The SMRC approach combined with the TGDR provides simultaneous built-in biomarker selection and classification rule estimation by optimizing a smoothed AUC as the objective function. This approach is computationally feasible for high-dimensional genomic data. For both Colon and Estrogen datasets, the estimation procedure takes ∼4 min using our current implementation of the algorithm in R (R Development Core Team, 2005, www.R-project.org). We plan to also implement our method in C. It is expected that the same computation can be accomplished using C program in ∼30 s. We note that the proposed approach is not necessarily optimal for all datasets in terms of classification error, since AUC and classification error are two different measures of classification performance. For example for the Estrogen data, the LogitBoost and the classification tree in Dettling and Buhlmann (2003) can have smaller classification errors (although note again different sets of genes are used). We expect that the relative performances of different approaches are data dependent and there exists no single dominating approach. A systematic comparison of different approaches is of interest for future study.

The sigmoid function has been extensively used in machine learning and neural network studies (Gammerman, 1996). Gammerman (1996) noted that the empirical sigmoid objective function may have multiple local minima in neural network studies. Several *ad hoc* solutions have been proposed. For example the global minimum can be detected by varying starting values for gradient search. It appears that similar simple solutions are also

applicable in the present setting. Although our own experiences show that this usually does not pose a serious problem, it is worth further investigation.

Classification using the ROC technique for two-class problems can be extended to multi-class problems. In the most general case, we obtain the classification rule by maximizing the volume under ROC surface (VUS), which generalizes the AUC for two-sample classification (Mossman, 1999). Provost and Domingos (2003) proposed the average of AUC from one-versus-rest comparisons weighted by the class probabilities as the objective function. Our proposed smoothing and regularization methods can be applied to these objective functions with minor modifications. We plan to study the regularized ROC method for multi-class problems in a future report.

ACKNOWLEDGEMENTS

REFERENCES

Abrevaya,J. (1999) Computation of the maximum rank correlation estimator. *Econ. Lett.*, **62**, 279–285.

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

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

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–584.

Cui,X. *et al.* (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Bioinformatics*, **6**, 59–75.

Dettling,M. and Buhlmann,P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics*, **9**, 1061–1069.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for tumor classification based on microarray data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Friedman,J.H. and Popescu,B.E. (2004) Gradient directed regularization for linear regression and classification. *Technical report*. Department of Statistics, Stanford University, CA.

Gammerman,A. (1996) *Computational Learning and Probabilistic Reasoning*. Wiley, New York.

Ghosh,D. and Chinnaiyan,A.M. (2005) Classification and selection of biomarkers in genomic data using LASSO. *J. Biomed. Biotechnol.*, **2**, 147–154.

Gui,J. and Li,H. (2005) Threshold gradient descent method for censored data regression with applications in pharmacogenomics. In *Proceedings of PSB 2005*. http://helix-web.stanford.edu/psb05/#pharmacogenomics.

Han,A.K. (1987) Non-parametric analysis of a generalized regression model. *J. Econometrics*, **35**, 303–316.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning* Springer-Verlag, New York.

Horowitz,J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**, 505–531.

Ma,S., Kosorok,M.R. and Fine,J.P. (2005) Additive risk models for survival data with high dimensional covariates. *Biometrics* (in press).

Mossman,D. (1999) Three-way ROCs. *Med. Decis. Making*, **19**, 78–89.

Nguyen,D. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Pepe,M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, UK.

Pepe,M.S. *et al.* (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.*, **159**, 882–890.

Pepe,M.S. *et al.* (2005) Combining predictors for classification using the area under the ROC curve. *Biometrics* (in press).

Pochet,N. *et al.* (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, **17**, 3185–3195.

Provost,F. and Domingos,P. (2003) Tree induction for probability based rankings. *Mach. Learning*, **52**, 199–215.

R Development Core Team (2005) R: a language and environment for statistical computing. Vienna, Austria.

Spang,R., Blanchette,C., Zuzan,H., Marks,J., Nevins,J. and West,M. (2001) Prediction and uncertainty in the analysis of gene expression profiles. In *Proceedings of the German Conference on Bioinformatics GCB 2001*.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Wahba,G. (1990) Spline models for observational data. In *Proceedings of the CBMS-NSF Regional Conference Series in Applied Mathematics,* **59**, SIAM, Philadelphia.

West,M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11562–11467.