BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 15 2007, pages 1945–1951
doi:10.1093/bioinformatics/btm287

Gene expression

Logistic regression for disease classification using microarray data: model selection in a large *p* and small *n* case

J.G. Liao[1,*] and Khew-Voon Chin[2]

[1]Drexel University School of Public Health, Philadelphia, PA 19102 and [2]The University of Toledo,Toledo, OH 43614, USA

Received on September 10, 2006; revised on May 14, 2007; accepted on May 21, 2007

Advance Access publication May 31, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Logistic regression is a standard method for building prediction models for a binary outcome and has been extended for disease classification with microarray data by many authors. A feature (gene) selection step, however, must be added to penalized logistic modeling due to a large number of genes and a small number of subjects. Model selection for this two-step approach requires new statistical tools because prediction error estimation ignoring the feature selection step can be severely downward biased. Generic methods such as cross-validation and non-parametric bootstrap can be very ineffective due to the big variability in the prediction error estimate.

Results: We propose a parametric bootstrap model for more accurate estimation of the prediction error that is tailored to the microarray data by borrowing from the extensive research in identifying differentially expressed genes, especially the local false discovery rate. The proposed method provides guidance on the two critical issues in model selection: the number of genes to include in the model and the optimal shrinkage for the penalized logistic regression. We show that selecting more than 20 genes usually helps little in further reducing the prediction error. Application to Golub's leukemia data and our own cervical cancer data leads to highly accurate prediction models.

Availability: R library *GeneLogit* at http://geocities.com/jg_liao

Contact: jl544@drexel.edu

1 INTRODUCTION

DNA microarray is a new technology that measures the expression levels of thousands of genes simultaneously and has emerged as an important tool in biomedical research. Golub *et al.* (1999) show that microarray gene expression can be used to classify between acute myeloid leukemia (AML) and acute lymphocytic leukemia. Since then, disease classification using microarray data has been the focus of intensive research with the aim of providing more accurate diagnostic tools than what the traditional pathological method alone can provide. Gene expression can also be used to predict survival time, disease prognostics and response to treatment, all with important clinical implications.

The mathematical and statistical methodologies for building such classification models, from the classical statistical methods to machine learning theory to classification trees, are reviewed and compared by Dudoit *et al.* (2002), Lee *et al.* (2005) and Li *et al.* (2004). This article considers the logistic regression approach, a standard method for binary classification that has been extended for use in microarray data in Eilers *et al.* (2001), Fort and Lambert-Lacroix (2005), Nguyen and Rocke (2002), Shen and Tan (2005), Zhou *et al.* (2004), Zhu and Hastie (2005). Let y be an array's binary disease status (1 for cancer and 0 for normal as a general example) and let $\mathbf{x} = (x_1, \ldots, x_p)$ be the expression vector, where x_j is the expression level of the jth gene. A logistic prediction model for $\pi(\mathbf{x})$, the probability of $y=1$ given $\mathbf{x}$, is constructed from a training dataset, which can then be applied to the gene expression of a new array to estimate its cancerous probability. Building a logistic prediction model using microarray data, however, is fundamentally different from standard logistic modeling because the number of genes (predictors) p can be thousands while the number of arrays (subjects) n is usually no more than 100. A popular approach is to combine a gene selection step with penalized likelihood inference (Eilers *et al.*, 2001; Shen and Tan, 2005; and Zhu and Hastie, 2004). Step 1, called feature selection, selects a subset of genes to include in the logistic regression. For ease of exposition, we will focus on the method of selecting the q most univariately significant genes (Dudoit *et al.*, 2002) and let $x_j^*, j = 1, \ldots, q$, be the expression of the jth selected gene. Let y_i, $i=1, \ldots, n$, be the binary disease status of the ith array in the training dataset and let $\mathrm{x}_i=(x_{i1}, \ldots, x_{ip})$ be its gene expression vector. Step 2 fits the logistic model

$$\mathrm{logit}\{\pi(x)\} = \beta_0 + \sum_{j=1}^{q} \beta_j x_j^* \quad (1)$$

by maximizing the penalized log-likelihood

$$l(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)\} - \frac{1}{2\tau^2}\|\beta\|^2, \quad (2)$$

where $\pi_i = \pi(\mathbf{x}_i)$ as given by model (1), $|\beta|$ is the Euclidean length of $\beta=(\beta_1, \ldots, \beta_q)$, and $\tau \in (0, \infty)$ is the shrinkage

*To whom correspondence should be addressed.

parameter that controls the degree of shrinkage of β toward 0 (Cessie and Van Houwelingen, 1990; Van Houwelingen, 2001). There are two key unresolved model selection issues here. The first is how to choose the number of genes q in Step 1. A smaller q makes the prediction model (1) easier and less costly to use but may lead to a larger prediction error. The second is how to find the optimal shrinkage parameter τ for a chosen q. To address these issues, we need to be able to estimate the prediction error of a model building strategy. Ambroise and McLachlan (2002) and Simon *et al.* (2003) show that methods in earlier publications that ignored the feature selection step in the evaluation can severely underestimate the prediction error. To incorporate both Steps 1 and 2 for valid assessment, they use the generic cross-validation and non-parametric bootstrap. Recently, however, Braga-Nato and Dougherty (2004) demonstrated that the prediction error estimation using cross-validation can be too variable to be useful. Efron (2004), in a significant theoretical advance, shows that cross-validation and non-parametric bootstrap, while broadly applicable, pay a substantial price in terms of decreased estimating efficiency, which is especially acute for a large p and small n problem. The parametric bootstrap method, based on models tailored to the specific problem, can offer substantially better accuracy if the model is justified.

This article proposes and develops a parametric bootstrap method for more accurate and reliable estimation of the prediction error for the two-step procedure of building a logistic model, which provides the critically needed guidance on the choice of q and τ. For any given q, our method finds the optimal τ and the corresponding prediction error. We show that including $q = 20$ genes in the model is usually sufficient as additional genes help little in further reducing the prediction error. Application to Golub's leukemia data (Golub *et al.*, 1999) and our own breast cancer data (Wong *et al.*, 2003) leads to highly accurate prediction models. A carefully crafted R library, *GeneLogit*, is supplied on our web-site (http://geocities.com/jg_liao) that can be readily used for data analysis.

2 PARAMETRIC BOOTSTRAP MODEL SELECTION

2.1 A parametric bootstrap model

The parametric bootstrap (Efron and Tibshirani, 1993), unlike cross-validation and non-parametric bootstrap, requires a more detailed model for the underlying process that generated data $\mathbf{y} = (y_1, \ldots, y_n)$. We now propose such a model in a hierarchical form.

Stage 1: given the gene expression vector $\mathbf{x}_i$, $y_i \sim$ Bernoulli (π_i) with

$$\text{logit}(\pi_i) = b_0 + \sum_{j=1}^{p} b_j x_{ij}, \; i = 1, \ldots, n, \tag{3}$$

where $b_1, \ldots, b_p$ are the regression coefficients to be further modeled in Stage 2 and b_0 is the intercept. Note that Equation (3) is different from Equation (1) in that (3) is the assumed true model for π_i while (1) is the working model for building a prediction rule. In Section 2.1, $\pi_1, \ldots, \pi_n$ are always as given in (3). Usually, only a small percentage of genes from (3) are needed in (1) because the expression levels of many genes are collinear.

Let n_1 be the total number of cancer arrays and $n_0 = n - n_1$ be the total number of normal arrays in the training sample. It can often be more appropriate to model $\mathbf{y}$ as drawn conditional on $y_1 + \ldots, + y_n = n_1$ because n_1 and n_0 are fixed by design (see Section 4 for more discussion). Let

$$\Pr(\mathbf{y}|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where π=$(\pi_1, \ldots, \pi_n)$. Let $\mathbf{u} = (u_1, \ldots, u_n)$, where each u_i is either 0 or 1, and let $S = \{\mathbf{u} : \sum_{i=1}^{n} u_i = n_1\}$. The conditional probability of $\mathbf{y}$ given $y_1 + \ldots, + y_n = n_1$ is then

$$\frac{\Pr(\mathbf{y}|\pi)}{\sum_{\mathbf{u}\in S} \Pr(\mathbf{u}|\pi)}, \tag{4}$$

which depends on $b_1, \ldots, b_p$ but not on intercept b_0 and is often the likelihood of choice in statistical inference when n_1 is fixed by design or when the intercept b_0 is considered a nuisance parameter (Agresti, 2002, McCullagh and Nelder, 1989; Chapter 6.7.1). We will therefore use this conditional distribution for the inference and simulation subsequently.

Stage 2: to develop a model for coefficients $b_1, \ldots, b_p$ in (3), let $M = \{j : b_j \neq 0\}$ be the subset of genes with a non-zero regression coefficient. Rewrite (3) as

$$\text{logit}(\pi_i) = b_0 + \sum_{j\in M} s_j |b_j| x_{ij}, \tag{5}$$

where $s_j = |b_j| / b_j$ is the sign of b_j for $j \in M$. Note that Equation (3) is ill-posed in that infinitely many solutions of $b_0, \ldots, b_p$ exist for any given π. To put a reasonable structure on b_j, we shall seek insight from a formulation of the gene expression vector $\mathbf{x}$ that leads to logistic regression (3). Assume that $\mathbf{x} \sim N(\mu_1, V)$ for a tissue drawn from the cancer group and $\mathbf{x} \sim N(\mu_0, V)$ for a tissue drawn from the normal group, where $\mu_1 = (\mu_{11}, \ldots, \mu_{1p})$ and $\mu_0 = (\mu_{01}, \ldots, \mu_{0p})$. This leads to logistic regression (3) with $(b_1, \ldots, b_p)^t = V^{-1}(\mu_1 - \mu_0)$ and intercept b_0 that depends on the relative sampling probabilities from the normal and cancer groups (Hastie, *et al.*, 2001). Assume further that the gene expression is independent across genes so that $V = \text{diag}(v_1, \ldots, v_p)$. We then have $b_j = (\mu_{1j} - \mu_{0j})/v_j$. Note that b_j is a multivariate regression coefficient in (3) while $\mu_{1j} - \mu_{0j}$ represents the marginal relationship between the jth gene's expression and the outcome y. Let H_j, j=1, ..., p, be the hypothesis that the jth's gene expression has no difference between the cancer and normal arrays, H_j^+ be the alternative hypothesis that expression is stronger in the cancer arrays and H_j^- be the hypothesis that the expression is stronger in the normal arrays. It follows that H_j, H_j^+ and H_j^- correspond to $b_j = 0$, $s_j = 1$ and $s_j = -1$, respectively. We can now borrow from the extensive research on identifying differentially expressed genes based on the false discovery rate (FDR) framework (Benjamini and Hochberg, 1995). Let z_j be the P-value from testing H_j that summarizes the strength of statistical evidence against H_j. Efron *et al.* (2001) and

Efron and Tibshirani (2002) model z_j, $j = 1, \ldots, p$, as generated from the mixture distribution

$$\eta_0 f_0(z) + (1 - \eta_0) f_1(z),$$

where η_0 is the proportion of genes that do not express differentially, f_0 is the uniform density on [0, 1] for P-values under the null hypothesis and f_1 is the density for P-values under the alternative. The local FDR, which quantifies the plausibility of individual H_j being true, is given by

$$\mathrm{fdr}_j = \Pr(H_j \text{is true} \,|z_j) = \frac{\eta_0}{\eta_0 + (1 - \eta_0) f_1(z_j)}, \qquad (6)$$

and can be estimated using the method either in Liao *et al.* (2004) or in Scheid and Spang (2005). To model the subset M, we shall generate it as a random set so that each j has probability $1 - \widehat{\mathrm{fdr}}_j$ of being in M independently for $j = 1, \ldots, p$. To determine a reasonable value for s_j, for $j \in M$, or to choose between H_j^+ and H_j^- after H_j is rejected, note that it is common practice to conclude that the direction of difference in the population is the same as what it is in the sample when a two-sided null hypothesis is rejected (Leventhal and Huynh, 1996). We will thus assign, for $j \in M$, $s_j = 1$ or $s_j = -1$ depending on the direction of the gene's expression in the training dataset. Because of the adjustment for multiple comparisons, a gene is only usually included in M when the P-value z_j is much smaller than the usual cut point 0.05. This way of modeling $b_j \neq 0$ and assigning s_j is based on the marginal relationship between the outcome y and the jth gene's expression level. We show that it is justified by assuming normal distribution for gene expression vector $\mathbf{x}$ and by assuming independent expression across genes within the normal group and within the cancer group. While this is somewhat a strong assumption, we think it provides a useful approximation to an otherwise intractable problem. Similar approach is adopted in Barbieri and Berger (2002) and Ishwaran and Rao (2005) where variables in a multivariate model are selected based on their individual performance instead of their joint posterior distribution in Bayesian analysis of large p and small n data. Further research is needed on the best ways to model b_j for a group of highly correlated genes.

To complete the specification for b_j, we shall model $|b_j|$, $j \in M$, as independent random effects from normal distribution $N(0, \theta_0^2)$ truncated on $(0, \infty)$. The variance component θ_0, to be estimated from the data, quantifies the size of $|b_j|$. This follows the established statistical tradition of modeling parameters of similar characteristics as random effects (Laird and Ware, 1982) and also naturally motivates penalized likelihood (2) (Cessie and Van Houwelingen, 1990; Van Houwelingen, 2001).

It is easy to draw a bootstrap sample $y^* = (y_1^*, \ldots, y_n^*)$ from the proposed two-stage model. To do this, first get an estimate of the local FDR as $\widehat{\mathrm{fdr}}_j$. Obtain an estimate of θ_0 *as* $\hat{\theta}_0$ by maximizing (7) as discussed subsequently. We can then draw $\mathbf{y}^*$ as follows.

Step 1: generate M so that each j has probability of $1 - \widehat{\mathrm{fdr}}_j$ of being in M. Assign, for $j \in M$, $s_j = 1$ if the gene's expression is stronger in cancer arrays for the training dataset and $s_j = -1$ otherwise. Draw $|b_j|$ from $N(0, \hat{\theta}_0^2)$ truncated on $(0, \infty)$ for $j \in M$. Model (3) or (5) is now specified except the intercept b_0.

Step 2: draw $\mathbf{y}^*$ from the conditional logistic distribution (4) with π given by model (5) as generated in Step 1. Note that both Steps 1 and 2 need to be performed for each bootstrap sample $\mathbf{y}^*$ as Step 1 models the uncertainty in regression coefficients in (3) or (5).

Finally, to estimate the variance component θ_0, let $\tilde{\mathbf{y}}$ be a bootstrap sample generated with $|b_j|$ in Step 1 drawn from the truncated $N(0, \theta^2)$ instead. Let $h(\theta)$ be the probability of $\tilde{\mathbf{y}} = \mathbf{y}$, where the probability incorporates both Steps 1 and 2 in drawing $\tilde{\mathbf{y}}$. Our estimate $\hat{\theta}_0$ maximizes

$$\theta^{-1/3} h(\theta), \qquad (7)$$

where the factor $\theta^{-1/3}$ is added to the likelihood function $h(\theta)$ to improve the stability of the maximization and can be seen as the kernel of a prior density on, say $[10^{-6}, 1]$, that mildly favors a smaller θ.

2.2 Estimation of the prediction error

With what is proposed in Section 2.1, the prediction error of a model building procedure can be easily evaluated using parametric bootstrap. For ease of exposition, we will denote the two-step procedure for building a logistic model in Section 1 by logistic (q, τ), where q is the number of the most significant genes included in working model (1) and τ is the shrinkage parameter in penalized likelihood (2). For any given pair (q, τ), the prediction error of procedure logistic (q, τ) can be estimated as follows:

Step 1: draw a bootstrap sample $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ as described in Section 2.1. Note that the intercept b_0 in (5) was left unspecified there but its value is needed here. Given the generated M, s_j and $|b_j|$, we will choose b_0 so that (5) satisfies $\pi_1 +, \ldots, + \pi_n = n_1$, which is the maximum likelihood estimate of b_0. The $\pi_1, \ldots, \pi_n$ are now completely specified in (5).

Step 2: apply the two-step procedure logistic (q, τ) to bootstrap sample $\mathbf{y}^*$. To be more specific, first test the gene expression difference between arrays with $y_i^* = 1$ and arrays with $y_i^* = 0$ for each of the p genes and select the q most significant genes. Then fit logistic regression (1) with y in penalized likelihood (2) replaced by $\mathbf{y}^*$. Let $\hat{\pi}^* = (\hat{\pi}_1^*, \ldots, \hat{\pi}_n^*)$ be the estimated $\pi_1, \ldots, \pi_n$ from the resulting model.

Step 3: compute the expected Brier score

$$n^{-1} \sum_{i=1}^{n} \left\{ (\hat{\pi}_i^* - \pi_i)^2 + \pi_i (1 - \pi_i) \right\}, \qquad (8)$$

which is smaller if the estimate $\hat{\pi}^*$ is closer to the true π. To derive (8), let $\mathbf{y}' = (y_1', \ldots, y_n')$ be n independent Bernoulli trials with each y_i' having success probability π_i. The Brier score (Brier, 1950) of using $\hat{\pi}^*$ to predict $\mathbf{y}'$ is

$$\text{Brier score} = n^{-1} \sum_{i=1}^{n} (\hat{\pi}_i^* - y_i')^2.$$

Formula (8) is the expected Brier score over $\mathbf{y}'$.

Repeat Steps 1–3 here a large number of times (10 000 times for the two examples below) and the average of (8) over

these replications serves as an estimate of the prediction error of procedure logistic (q, τ).

As mentioned earlier, the closer the bootstrap estimate $\hat{\pi}^{\star}$ is to the underlying true π, the smaller the expected Brier score (8) is. The true π from Step 1, however, is a random quantity determined by the generated values of M, s_j and $|b_j|$, which models our uncertainty about π. The prediction error estimate of procedure logistic (q, τ) averages (8) over π. Our parametric bootstrap method can therefore be alternatively motivated from the perspective of Bayesian model averaging (Hoeting *et al*., 1999).

2.3 Software implementation

Our proposed method is implemented in software library *GeneLogit* that runs on the open source R environment (R Development Core Team, 2006) and is available at http://geocities.com/jg_liao For our implementation, the local FDR estimator in Liao *et al*. (2004) is used. The conditional probability (4) is generally computationally onerous because the probability of $y_1 +, \ldots, + y_n = n_1$ in the denominator may involve the sum of a large number of terms. To simplify the computation, normal approximation $\Phi(n_1 + 0.5) - \Phi(n_1 - 0.5)$ is used, where Φ is the cumulative distribution of normal distribution with mean $\pi_1 +, \ldots, + \pi_n$ and variance $\pi_1(1-\pi_1) +, \ldots, + \pi_n(1-\pi_n)$. The free parameter b_0 in (3) is chosen so that $\pi_1 +, \ldots, + \pi_n = n_1$ for more accurate approximation, which works well when $n > 30$ and n_1 is not too close to 0 or n. The likelihood $h(\theta)$ in (7) is computed by averaging conditional probability (4) over 5000 simulated $b_1, \ldots, b_p$, where $b_1, \ldots, b_p$ are generated as in Step 1 in Section 2.1 except $|b_j|$ are drawn from $N(0, \theta^2)$ truncated on $(0, \infty)$. The resulting library *GeneLogit* is easy to use but still computationally intensive (~20 hours of total computing time on today's dual core CPU from Intel or AMD for either the leukemia or the cervical cancer dataset subsequently). As input, it requires the disease outcome **y** as a vector and the gene expression as an $n \times p$ matrix. No missing values are allowed. The gene expression matrix needs to be standardized so that each column (the expression of a gene across all subjects) has variance of 1. There is no need, however, to center the gene expression to 0 as it does not impact on the regression coefficients $b_1, \ldots, b_p$. The program seems numerically stable as same result is obtained from many different runs. We now illustrate the use of the library in two datasets.

3 TWO EXAMPLES

3.1 Golub leukemia data: classification between ALL and AML

Golub *et al*. (1999) use gene expression to classify between acute lymphoblastic leukemia (ALL) and AML. Their data have since been analyzed by many authors using different classification methodologies. The training dataset consists of 27 ALL and 11 AML subjects and the test dataset consists of 20 ALL and 14 AML subjects. The expression of 7129 genes were originally measured. We applied the pre-processing procedures in Dudoit *et al*. (2002) to filter out genes that do not exhibit significant variation across the samples, followed by thresholding and the logarithmic transformation. A total of 3051 genes remain after the pre-processing. Our proposed method is applied to the training dataset ($n = 38$ and $p = 3051$) in the following steps using *GeneLogit* library:

Table 1. The optimal τ and the prediction error for different q, leukemia data

q	1	10	20	50	100
Optimal τ	4.542	1.041	0.628	0.524	0.487
Prediction error	0.0191	0.0167	0.0152	0.0142	0.0141

(1) Run function *localFDR* to estimate fdr_j and assign s_j.

(2) Run function *model.estimation* to estimate θ_0 by maximizing (7), which gives $\hat{\theta}_0 = 0.0123$.

(3) Use function *bootstrap.prediction* to find the optimal τ and the corresponding prediction error for different values of q. The result for $q = 1$, 10, 20, 50 and 100 is given in Table 1. For example, the optimal τ for $q = 1$ is found to be 4.542 with a corresponding prediction error estimated to be 0.0191 and the optimal τ for $q = 20$ is 0.628 with a corresponding prediction error of 0.0152. A larger q leads to a smaller optimal τ (more shrinkage) and smaller prediction error. Increasing q beyond 20, however, does not help much in further reducing the prediction error. Note that the estimated prediction error is the average over 10 000 expected Brier scores defined in Equation (8) from 10 000 bootstrapped models. For $q = 20$ and $\tau = 0.628$, the 5th and 95th percentiles of these expected Brier score are 0.0116 and 0.0381, respectively.

(4) Based on the result in Table 1, we choose procedure logistic ($q = 20$, $\tau = 0.628$) to build our prediction model by running function *pena.logit*. The 20 selected genes and the penalized logistic regression coefficients are given in Table 2. Note that the regression coefficients are all close to 0 due to the shrinkage effect.

To assess the prediction capacity of the model in Table 2 for independent samples, we apply it to Golub's test dataset of 20 ALL and 14 AML subjects. For each of the 34 subjects in the test dataset, we compute $\hat{\pi}_i$, the estimated probability of the ith subject being ALL, by applying the prediction model to the array's expression vector. Note, however, the number of cases over the number of normal subjects is 27/11 for the training dataset and 20/14 for the test dataset. The estimated intercept -0.278 needs to be adjusted to $-0.278 - \log(27/11) + \log(20/14)$ to account for the different sampling ratios (McCullagh and Nelder, 1989, Chapter 4.3.3). The $\hat{\pi}_i$ and the true disease status y_i (1 for ALL and 0 for AML) are given in Table 3 for $i = 1, \ldots, 34$. We see that $\hat{\pi}_i > 0.5$ for every subject with $y_i = 1$ and $\hat{\pi}_i < 0.5$ for every subject with $y_i = 0$ except one with $\hat{\pi}_i = 0.506$. The Brier Score for predicting $y_1, \ldots, y_{34}$ using $\hat{\pi}_1, \ldots, \hat{\pi}_{34}$ is 0.015. As comparison, Nguyen and Rocke (2002) report 1–3 classification errors; Lee *et al*.

Table 2. Logistic prediction model using procedure logistic($q = 20$, $\tau = 0.629$), leukemia data

Gene	Intercept	M55150.at	X95735.at	M27891.at	HG1612-HT1612.at	M16038.at
coefficient	−0.278	−0.261	−0.240	−0.265	0.223	−0.252
Gene	M27783.s.at	M31523.at	Z15115.at	D88422.at	X51521.at	M21551.rna1.at
coefficient	−0.165	0.224	0.260	−0.177	0.257	−0.240
Gene	U22376.cds2.s.at	M23197.at	X62320.at	Y12670.at	U50136.rna1.at	M63138.at
coefficient	0.263	−0.268	−0.177	−0.365	−0.246	−0.256
Gene	L09209.s.at	M31166.at	M31166.at			
coefficient	−0.237	−0.268	−0.299			

Table 3. The true disease status y_i and the estimated π_i for the 34 arrays in Golub's test dataset

y_i	1	1	1	1	1	1	1	1	1	1	1	1
$\hat{\pi}_i$	0.991	0.914	0.983	0.845	0.991	0.963	0.995	0.965	0.993	0.979	0.991	0.987
y_i	1	1	1	1	1	1	1	1	0	0	0	0
$\hat{\pi}_i$	0.939	0.960	0.986	0.980	0.728	0.903	0.961	0.986	0.012	0.034	0.008	0.004
y_i	0	0	0	0	0	0	0	0	0	0		
$\hat{\pi}_i$	0.046	0.027	0.036	0.244	0.212	0.122	0.506	0.023	0.035	0.021		

(2003) and Zhou *et al*. (2004) report one classification error and Yeung *et al*. (2005) have two classification errors. The result reported here can be reproduced by running the R code included in our *GeneLogit* library.

3.2 Classification between cervical cancer and normal tissues

Wong *et al*. (2003) study the gene expression of 26 cervical cancer tissues and 9 normal cervical tissues. The data is also analyzed in Liao *et al*. (2004) in the context of estimating the local FDR. We now use the data to build a logistic prediction model for classifying between cervical cancer and normal tissues. We use the 3670 genes, among 10 692 assessed, that have complete data for all the 35 subjects. Applying our *GeneLogit* library using the same steps as for the Golub's data, we obtain $\hat{\theta}_0 = 0.00901$. The optimal τ and the corresponding prediction error for $q = 1$, 10, 20, 50 and 100 are given in Table 4. Again, the optimal τ decreases as q increases and choosing q beyond 20 does not help much in further reducing the prediction error. We thus choose procedure logistic ($q = 20$, $\tau = 0.384$) to build our prediction model with result given in Table 5. For this dataset, however, there is no separate test dataset to reliably assess the prediction error for independent samples. To overcome this problem, we shall borrow idea from cross-validation. For $i = 1, \ldots, 35$, we apply procedure logistic ($q = 20$, $\tau = 0.384$) to the cervical dataset with data from the ith subject removed. The resulted logistic model is then applied to the ith tissue's gene expression to compute the estimate $\hat{\pi}_i^{-i}$. The result is given in Table 6. We see that all the cervical cancer tissues ($y_i = 1$) have $\hat{\pi}_i^{-i} > 0.5$ and all the normal tissues ($y_i = 0$) have $\hat{\pi}_i^{-i} < 0.5$. The Brier score for predicting $y_1, \ldots, y_{35}$ using $\hat{\pi}_i^{-i}, i = 1, \ldots, 35$, is 0.023. Note that y_i does not contribute to the value of $\hat{\pi}_i^{-i}$. The high agreement between y_i and $\hat{\pi}_i^{-i}$ in Table 6 for all 35 subjects indicates that the procedure logistic ($q = 20$, $\tau = 0.338$) is a good choice.

Table 4. The optimal τ and the prediction error for different q, cervical cancer

q	1	10	20	50	100
Optimal τ	2.73	0.795	0.384	0.279	0.207
Prediction error	0.082	0.059	0.057	0.054	0.053

4 CONCLUSION AND DISCUSSION

Building a logistic prediction model using microarray data poses considerable technical challenge because of the larger p and small n. Infinitely number of solutions of b_0, $b_1, \ldots, b_p$ exist for any given $\pi_1, \ldots, \pi_n$ for logistic model (3). Consequently, any observed data $y_1, \ldots, y_n$ is subject to infinitely many interpretations. Additional structure on b_0, $b_1, \ldots, b_p$ is required for effective data analysis. We propose such a structure in Section 2.1 by writing (3) in the form of (5) and by borrowing from the extensive research on FDR. In our model, the probability of $b_j \neq 0$ or $j \in M$ is taken to be $1 - \widehat{\text{fdr}}_j$ with the sign $s_j, j \in M$, obtained from the

Table 5. Logistic prediction model using procedure logistic (q=20, τ =0.384), cervical cancer

Gene	Intercept	AA598513	N73975	AA160783	R41779	AI220203
coefficient	7.458	0.174	−0.269	−0.197	0.168	0.200
Gene	H54629	AA148737	AA450265	AA669557	H88540	AI313031
coefficient	0.190	0.158	0.160	0.188	0.187	0.182
Gene	AA027161	AA488627	AA865342	AA487488	AI368402	H65335
coefficient	0.212	0.153	0.153	0.139	0.210	0.193
Gene	AA447797	AA452556	R85257			
coefficient	0.173	−0.177	0.137			

Table 6. The true disease status y_i and the estimated π_i for the 38 arrays in the cervical cancer dataset

y_i	1	1	1	0	1	1	1	1	1	1	1	
$\hat{\pi}_i^{-i}$	0.959	0.752	0.967	0.188	0.974	0.912	0.693	0.983	0.982	0.977	0.918	
y_i	1	0	1	1	0	1	1	1	1	0	1	0
$\hat{\pi}_i^{-i}$	0.975	0.222	0.994	0.992	0.261	0.953	0.973	0.936	0.988	0.0376	0.976	0.363
y_i	1	0	0	1	0	0	1	1	1	1	1	
$\hat{\pi}_i^{-i}$	0.946	0.122	0.173	0.658	0.272	0.280	0.965	0.876	0.987	0.967	0.975	

direction of the jth gene expression in the training dataset and $|b_j|$ modeled as random effects. The prediction error of procedure logistic (q, τ) can then be estimated using parametric bootstrap to guide the choice of q and τ. Application of the proposed method to the leukemia and cervical cancer datasets results in excellent prediction models. Biomedical researchers, we interacted with, found the method intuitive and easy to understand.

We now briefly discuss a few issues. First, the number of cancer arrays n_1 and the number of normal arrays n_0 in a microarray study are often chosen to be of similar size to increase statistical efficiency even though the underlying normal population can be much larger than the cancer population. This is similar to the case-control design in epidemiology in which individuals in the disease population are often sampled for inclusion in study in greater proportion than subjects in the control population. For such design, the intercept β_0 in (1) depends on the ratio of the sampling proportions but $\beta_1, \ldots, \beta_q$ do not (Agresti, 2002; McCullagh and Nelder, 1989, Chapter 6.7.1). In applying the logistic prediction model to new arrays, it is prudent to find out if the same ratio of sampling proportions is used as in the training dataset. The relatively cancerous risk, however, can be determined from only $\beta_1, \ldots, \beta_q$. Second, we have focused on the model-building procedure logistic(q, τ) in which the q univariately most significantly genes are included in the logistic model. Other feature selection methods can also be used. One may consider, e.g. to include the q jointly most significant genes. Our proposed bootstrap method can be used in the same way in evaluating its prediction error. Computationally, however, it is much more intensive to find the q jointly most significant genes. Third Yeung, *et al*. (2005) use Bayesian model averaging for building microarray classification models. As discussed in Section 2.2, our proposed bootstrap method can also be motivated from the same perspective. But Yeung *et al*. method is, in our opinion, a somewhat *ad hoc* adaptation of standard Bayesian model averaging to the large p and small n microarray data while we have developed a coherent model specifically tailored to such data. Indeed, they report more misclassified arrays for Golub's leukemia data.

ACKNOWLEDGEMENTS

Part of the first author's work was conducted when he was Biostatistician in the Cancer Center of New Jersey. The research was supported by NCI grant 2P30 CA 72720-04. Dr Yong Lin helped with computation. The authors thank the two reviewers whose comments led to substantial improvement of the manuscript.

Confilict of Interest : none declared.

REFERENCES

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

Agresti,A. (2002) *Categorical Data Analysis*. 2nd edn. Wiley, New York.

Barbieri,M.M. and Berger,J.O. (2004) Optimal predictive model selection. *Ann. Stat.*, **32**, 870–897.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **B57**, 289–300.

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

Brier,G.W. (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78**, 1–3.

Cessie,S.L. and Van Houwelingen,H.C. (1990) Ridge estimators in logistic regression. *Appl. Stat*., **41**, 191–201.

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

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–536.

Efron,B. (2004) The estimation of prediction error: covariance penalties and cross-Validation. *J. Am. Stati. Associ*., **99**, 619–632.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapmann and Hall/CRC, London.

Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol*., **23**, 70–86.

Efron,B. *et al*. (2001) Empirical Bayes analysis of microarray experiment. *J. Am. Stati. Associ*., **96**, 1151–1160.

Eilers,P.H. *et al*. (2001) Classification of microarray data with penalized logistic regression. *In Proceedings of SPIE Volume 4266: Progress in Biomedical optics and Imaging*, **Vol 2**, 187–198.

Fort,G. and Lambert-Lacroix,S. (2005) Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**, 1104–1111.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning*. Springer, Berlin.

Hoeting,J. *et al*. (1999) Bayesian model Averaging. *Stat. Sci*., **14**, 382–401.

Laird,N.M. and Ware,J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lee,K.E. *et al*. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Lee,J.W. *et al*. (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. & Data Anal*., **48**, 869–885.

Leventhal,L. and Huynh,C. (1996) Directional decisions for two-tailed tests: power, error rates, and sample size. *Psychol. Methods*, **1**, 278–292.

Li,T. *et al*. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Liao,J.G. *et al*. (2004) A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**, 2694–2701.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman and Hall, New York.

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

R Development Core Team *R: a Language and Environment for Statistical Computing*. R. fondtation for satistical computing. Vienna, AustriaISBN 3-900051-00-3 http://www.R-project.org.

Scheid,S. and Spang,R. (2005) Twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics*, **21**, 2921–2922.

Shen,L. and Tan,E.C. (2005) Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Trans. Compu. Biol. Bioinform*., **2**, 166–175.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Nat. Cancer Inst*, **95**, 14–18.

Van Houwelingen,J.C. (2001) Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, **55**, 17–34.

Wong,Y.F. *et al*. (2003) Expression genomics of cervical cancer: molecular classification and prediction of radio-therapy response by DNA microarray. *Clini. Cancer Res*., **9**, 5486–5492.

Yeung,K.Y. *et al*. (2005) Bayesian model averaging: development of an improved multi-calss, gene selection and classification tool for microarray data. *Bioinformatics*, **21**, 2394–2402.

Zhu,J. and Hastie,T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427–443.

Zhou,X. *et al*. (2004) Cancer classification and prediction using logistic regression with Bayesian gene selection. *J. Biomed. Inform*., **37**, 249–259.