



How many samples are needed to build a classifier: a general sequential approach

Wenjiang J. Fu^{1,*}, Edward R. Dougherty^{2,3}, Bani Mallick¹ and Raymond J. Carroll¹

¹Department of Statistics, Texas A&M University, 447 Blocker Building, College Station, TX 77843, USA, ²Department of Electrical Engineering, Texas A&M University, 214 Zachry Engineering Center, College Station, TX 77840, USA and ³Department of Pathology, University of Texas MD Anderson Cancer Center, 1515 Holcombe, Houston, TX 77030, USA

Received on April 20, 2004; revised on July 5, 2004; accepted on July 28, 2004
Advance Access publication August 5, 2004

ABSTRACT

Motivation: The standard paradigm for a classifier design is to obtain a sample of feature-label pairs and then to apply a classification rule to derive a classifier from the sample data. Typically in laboratory situations the sample size is limited by cost, time or availability of sample material. Thus, an investigator may wish to consider a sequential approach in which there is a sufficient number of patients to train a classifier in order to make a sound decision for diagnosis while at the same time keeping the number of patients as small as possible to make the studies affordable.

Results: A sequential classification procedure is studied via the martingale central limit theorem. It updates the classification rule at each step and provides stopping criteria to ensure with a certain confidence that at stopping a future subject will have misclassification probability smaller than a predetermined threshold. Simulation studies and applications to microarray data analysis are provided. The procedure possesses several attractive properties: (1) it updates the classification rule sequentially and thus does not rely on distributions of primary measurements from other studies; (2) it assesses the stopping criteria at each sequential step and thus can substantially reduce cost via early stopping; and (3) it is not restricted to any particular classification rule and therefore applies to any parametric or non-parametric method, including feature selection or extraction.

Availability: R-code for the sequential stopping rule is available at <http://stat.tamu.edu/~wfu/microarray/sequential/R-code.html>

Contact: wfu@stat.tamu.edu

1 INTRODUCTION

The standard paradigm for a classifier design is to obtain a sample of feature-label pairs and then to apply a classification

rule to derive a classifier from the sample data. In some cases, where sample data can be cheaply obtained or generated via a mathematical model, it may be possible to estimate the sample size necessary to obtain a desired degree of design precision and then proceed to generate a sample of the desired size. Typically in laboratory situations the sample size is limited by cost, time or availability of sample material. This study has been motivated by gene-expression-based cancer classification, where sample mRNA may be severely limited. The sample mRNA may have to be obtained from available tissue taken from deceased patients and for which careful and extensive pathology has been undertaken, or it may be obtained in clinical studies, where recruiting patients can be very costly, with costs ranging from providing free medical care to long-term follow-up. Thus, an investigator may wish to consider a sequential approach in which there is a sufficient number of patients to train a classifier in order to make a sound decision for diagnosis while at the same time keeping the number of patients as small as possible to make the studies affordable.

Consider a study that sequentially obtains and categorizes sample tissue or recruits and diagnoses patients based on certain clinical conditions, the goal being to design a classifier to discriminate among the kinds, attributes or stages of a disease. Rather than apply a procedure involving a pre-determined fixed size sample, we will take a sequential approach. The diagnostic decision follows a learning process, where the decision criteria update with the recruiting process with the aim of making as few misclassifications as possible. The key issue to be addressed is the formulation of a stopping rule for the sequential classification.

2 STOPPING RULE FOR THE SEQUENTIAL PROCEDURE

To formulate the stopping rule, let us set up the relevant random variables and probabilities. Let $\mathcal{Y}_i = \{Y_1, \dots, Y_i\}$,

*To whom correspondence should be addressed.

where $i = 1, 2, \dots$, be a set of binary observations in a sequential recruitment of patients, either being diagnosed of disease of interest (1) or no disease (0). Patients are recruited independently from each other, and are not correlated, i.e. no observations within the same family or unit are observed. Each patient will have clinical features or expression levels g_i of a certain number of genes. Let Q_i be the indicator function that Y_i is misclassified based on \mathcal{Y}_{i-1} , i.e. $Q_i = 1$ if Y_i is misclassified or $Q_i = 0$ otherwise. And let $\pi_i = P(Q_i = 1 | \mathcal{Y}_{i-1})$ be the conditional misclassification probability. We are thus interested in the number of patients (N) we need to recruit before we can claim with certain confidence $(1 - \alpha)\%$ that the next patient to recruit will have a small probability π_N of being misclassified with $\pi_N \leq \varepsilon$ for a given small threshold $\varepsilon > 0$.

To derive the stopping rule, we make the following assumptions:

- (1) The conditional probability π_n is weakly monotonically decreasing, i.e. there exists a finite integer $p_0 > 0$ such that $\pi_{i+p} \leq \pi_i$ almost surely for all $p \geq p_0$ and $i \geq 1$. Thus π_n converges weakly to $\pi_\infty \geq 0$.
- (2) The limit probability $\pi_\infty > 0$, i.e. there is a positive probability that observations may be misclassified. This is generally true in applications since the class distributions are assumed to overlap on a region of positive probability, with π_∞ depending on the type of classifier considered, such as linear discriminant analysis (LDA), and on the scientific context of the problems. For a given classification problem with a non-zero Bayes error, the probability $\pi_\infty > 0$ regardless of the type of classifiers.

The stopping rule that we apply depends on the following theorem, which depends on the martingale central limit theorem, and is proven in the Appendix.

THEOREM. For level $0 < \alpha < 1$,

$$P\left(\pi_N \leq N^{-1} \sum_{i=1}^N Q_i + z_{1-\alpha} \hat{\kappa}_N / N^{1/2}\right) \rightarrow 1 - \alpha \text{ as } N \rightarrow \infty,$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of the standard Gaussian distribution $N(0, 1)$,

$$\hat{\kappa}_N = N^{-1} \sum_{i=1}^N \hat{\pi}_i (1 - \hat{\pi}_i), \text{ and } \hat{\pi}_i = i^{-1} \sum_{j=1}^i Q_j$$

is the estimator for the conditional probability π_i .

The theorem suggests that a conservative sample size N , for given threshold $0 < \varepsilon < 1$, satisfies

$$\varepsilon = N^{-1} \sum_{i=1}^N Q_i + z_{1-\alpha} \hat{\kappa}_N / N^{1/2}.$$

Thus

$$N = \left(\frac{z_{1-\alpha} \hat{\kappa}_N}{\varepsilon - N^{-1} \sum_{i=1}^N Q_i} \right)^2$$

with $\varepsilon > N^{-1} \sum_{i=1}^N Q_i$ and $\hat{\kappa}_N > 0$. In addition, to ensure a stopping of the sequential procedure when a sufficiently large number N_0 of consecutive perfect classifications occur, we set $N_0 = \log(\alpha) / \log(1 - \varepsilon)$. Therefore, the number of sequential recruitments N in the stopping rule satisfies

$$N \geq \min \left[\left(\frac{z_{1-\alpha} \hat{\kappa}_N}{\varepsilon - N^{-1} \sum_{i=1}^N Q_i} \right)^2, N_0 \right],$$

$\hat{\kappa}_N > 0$ and $0 < N^{-1} \sum_{i=1}^N Q_i < \varepsilon$.

Based on this stopping rule and a chosen value of ε , the number of samples and the designed classifier are determined by the following algorithm under the assumption that the maximum sample size is M :

- (1) Start with an initial sample S_0 and set the perfect-classification counter $N_0 = 0$.
- (2) At step i , train the classifier based on current data.
- (3) Recruit one subject and classify based on the clinical feature using the classifier.
- (4) Set $Q_i = 0$ and $N_0 = N_0 + 1$ if the classification is correct; otherwise, set $Q_i = 1$ and $N_0 = 0$.
- (5) Calculate $\hat{\kappa}_N$.
- (6) If the stopping rule is satisfied or $N = M$, then stop; otherwise, go back to Step 2 and repeat Steps 2–6.

While a sequence of consecutive perfect classifications indicates good performance of the classifier and induces early stopping, large Bayes errors, such as 30% or greater, make it difficult to train good classifiers and also ought to induce early stopping to save resources. This involves a hypothesis testing with null hypothesis $H_0 : \pi_\infty \leq 0.3$ versus the alternative $H_1 : \pi_\infty > 0.3$. Much work has been done in the area of hypothesis testing in sequential trials (see Emerson and Fleming, 1990; Lai, 1997 and references therein).

It is well-known that maximum-likelihood estimations in sequential studies are biased due to early stopping, and thus need to be corrected, (see Liu and Hall, 1999; Pinheiro and DeMets, 1997; Todd and Whitehead, 1996; Whitehead, 1986). In this paper, we focus on training classifiers rather than on estimating misclassification error, and thus refer readers to the above references for methods of error estimation and bias correction.

3 WEAK CONVERGENCE WITH LARGE SAMPLES

We consider the large sample behavior for the classifier built by the sequential procedure. Let C_n be the classifier trained

based on the observation \mathcal{Y}_n . C_n is a random function. Let Y be a new observation. By large sample theory, the weak convergence $C_n(Y) \rightarrow_p C_\infty(Y)$ holds for some classifier C_∞ . Thus

$$\pi_{n+1} = P(Q=1|\mathcal{Y}_n) = P_{C_n}(Q=1|\mathcal{Y}_n) \rightarrow_p P_{C_\infty}(Q=1|\mathcal{Y}_\infty),$$

where \mathcal{Y}_∞ is the sequence of observations \mathcal{Y}_m with m being infinitely large. Since C_∞ does not depend on the observations \mathcal{Y}_∞ , $\pi_n \rightarrow_p P_{C_\infty}(Q=1)$, i.e. the conditional probability of misclassification in the sequential procedure converges in probability to the limit probability for misclassification. It implies that although we set a threshold on the conditional probability in the sequential procedure, it is approximately equivalent to setting a threshold on the unconditional probability for sufficiently large number of sequential steps.

REMARKS

- (1) $P_{C_\infty}(Q=1)$ depends on the type of classifier predetermined, and may assume different values for different types of classifiers. It may be much larger than the misclassification rate of the optimal Bayes classifier.
- (2) In practice, if the Bayes classifier has a positive misclassification error, and no matter how small ε is chosen to be, the resulting classifier trained by the sequential procedure cannot perform better—that is, the misclassification error of C_N cannot be smaller than the Bayes error even if a smaller ε is selected. This will be demonstrated in our simulation studies.

4 SIMULATION STUDY

To demonstrate the basic behavior of sequential procedure, we compare the misclassification error of the trained classifier with Bayes error in the simple case of a single feature and two Gaussian class-conditional distributions $N(0, 1)$ and $N(\Delta, 1)$, $\Delta > 0$, where $N(0, 1)$ and $N(\Delta, 1)$ correspond to the labels 0 and 1, respectively. The misclassification error of the designed classifier is its true error computed from the model. In this case, the Bayes classifier corresponds to the optimal linear classifier and we use LDA. The sequential procedure starts with 10 random samples, 5 from each class, proceeds according to the algorithm with a maximum of 90 sequential steps, and yields a classifier based on a sample size of N .

To examine classifier performance, we drew random samples of 10 000 data points, 5000 from $N(0, 1)$ and 5000 from $N(\Delta, 1)$, and tested the LDA classifier to estimate its cutoff value λ . Thus a data point is categorized as class 0 if less than λ , or class 1 otherwise. We repeated the drawing of random samples 50 times to obtain accurate estimation of λ . The error of the LDA classifier was then calculated with $e(C_N) = [1 - \Phi(\lambda) + \Phi(\lambda - \Delta)]/2$ for each fixed value Δ . This sequential training and testing procedure was repeated 1000 times for each fixed pair of (Δ, ε) to compare the LDA error with the Bayes error, which was calculated with

$[1 - \Phi(\Delta/2)]$. The sample mean misclassification rate and sample SD of 1000 simulation runs with a random sequence of observations in each run were computed and the minimum, maximum and mean numbers of sequential steps in the procedure were also recorded.

Table 1 reports the mean misclassification rate, SD of the rates over 1000 simulation runs, the minimum, maximum and mean numbers of sequential steps, and the SD in training the classifiers for each Δ and ε . The means and SD were computed with the sample mean and sample variances. The maximum of sequential steps is 90. The misclassification rates of the classifiers trained with the sequential procedure are very close to the Bayes errors. Specifying a smaller threshold ε increases the number of sequential steps, sometimes substantially, but the overall misclassification rate is only lowered slightly. The most gain in Table 1 is for $\Delta = 2.3$, with the misclassification rate decreasing by 0.0084 from $\varepsilon = 0.2$ to 0.05, a 6% reduction in error at the cost of increasing the number of steps from 16 to 83 on the average. Therefore, specifying a very small threshold ε is not recommended in practice when patient recruitment is costly.

Figure 1 shows the histogram plots of the misclassification errors over 1000 simulation runs by Δ and ε . The arrow in the bottom left corner of each plot indicates the Bayes error for given Δ , which is the minimum value in each histogram. Although long tails to the right were observed, the tail probabilities are very small and the error estimates are in general very close to the Bayes error. It indicates that the LDA classifier at the stopping performs well with only slight bias.

Table 2 shows analogous results for the maximum number of sequential steps being 40. It shows that the mean misclassification error increases with smaller maximum number of sequential steps allowed compared to Table 1. Again, the number of sequential steps is reduced substantially while the increase in mean misclassification error varies, mostly for large Bayes error corresponding to $\Delta = 1$, and barely for small Bayes error corresponding to $\Delta = 4$.

Tables 1 and 2 indicate that setting a smaller threshold ε or increasing the maximum number of steps allowed in the sequential procedure may achieve smaller misclassification error, but the reduction may be slight relative to the increase in the number of required patients.

5 APPLICATION TO MICROARRAY DATA

We apply the sequential procedure to two studies of microarray data, one on breast-cancer patient prognosis, and the other on breast-tumor characterization. van't Veer *et al.* (2002) and van de Vijver *et al.* (2002) have reported studies of gene expression profiling on breast-cancer patient prognosis based on 295 samples with 70 selected genes. Perou *et al.* (2000) have considered the molecular portrait of human breast tumors based on 84 samples with expressions of 8102 genes. We have chosen these two datasets because the large sample sizes facilitate

Table 1. Misclassification error, the SD, minimum, maximum and mean (\bar{N}) numbers of sequential steps and the SD in training LDA classifier by population mean Δ and stopping threshold ε for fixed $\alpha = 0.05$ and maximum sequential steps 90

| Δ | Bayes error ^a | | ε | | | |
|----------|--------------------------|-------------------|-----------------|-----------------|-----------------|-----------------|
| | | | 0.05 | 0.1 | 0.15 | 0.2 |
| 1 | 0.3085 | Err (SD) | 0.3134 (0.0068) | 0.3143 (0.0074) | 0.3141 (0.0077) | 0.3147 (0.0083) |
| | | (min, max) | (26, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (90, 2) | (87, 15) | (80, 26) | (70, 34) |
| 1.3 | 0.2578 | Err (SD) | 0.2617 (0.0056) | 0.2617 (0.0058) | 0.2634 (0.0100) | 0.2651 (0.0119) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (90, 5) | (84, 20) | (72, 32) | (57, 37) |
| 1.5 | 0.2266 | Err (SD) | 0.2301 (0.0051) | 0.2304 (0.0058) | 0.2323 (0.0100) | 0.2340 (0.0119) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (89, 8) | (80, 25) | (67, 35) | (45, 37) |
| 2 | 0.1587 | Err (SD) | 0.1612 (0.0039) | 0.1629 (0.0071) | 0.1654 (0.0106) | 0.1681 (0.0146) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (87, 14) | (67, 34) | (43, 35) | (25, 25) |
| 2.3 | 0.1251 | Err (SD) | 0.1277 (0.0042) | 0.1300 (0.0089) | 0.1329 (0.0137) | 0.1361 (0.0167) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (83, 20) | (54, 35) | (31, 29) | (16, 16) |
| 2.5 | 0.1056 | Err (SD) | 0.1087 (0.0057) | 0.1113 (0.0103) | 0.1134 (0.0145) | 0.1164 (0.0182) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 90) |
| | | (\bar{N} , SD) | (78, 25) | (48, 34) | (25, 23) | (15, 14) |
| 3 | 0.0668 | Err (SD) | 0.0701 (0.0068) | 0.0732 (0.0103) | 0.0752 (0.0138) | 0.0770 (0.0160) |
| | | (min, max) | (21, 90) | (11, 90) | (8, 90) | (6, 56) |
| | | (\bar{N} , SD) | (65, 29) | (29, 22) | (17, 12) | (12, 6) |
| 4 | 0.0228 | Err (SD) | 0.0263 (0.0060) | 0.0290 (0.0130) | 0.0300 (0.0132) | 0.0325 (0.0205) |
| | | (min, max) | (21, 90) | (11, 85) | (8, 37) | (6, 28) |
| | | (\bar{N} , SD) | (45, 20) | (24, 9) | (16, 4) | (12, 3) |

^aBayes error between the two populations Normal (0, 1) and Normal (Δ , 1) was calculated with $e_B = 1 - \Phi(\Delta/2)$, where Φ is the cumulative distribution function of the distribution Normal (0, 1).

accurate estimation of the misclassification errors of the trained classifiers. The breast tumor study has five classes of tissues: luminal-like ER + tumors, ERBB2 + tumors, normal breast, breast cell lines and basal-like tumors. For simplicity, we pool them further and form two classes, 51 samples of tumor-like tissues (including luminal-like ER+ tumors, ERBB2 + tumors and basal-like tumors) and 33 samples of non-tumor tissues (including normal breast and breast cell lines).

To initialize the sequential procedure, we choose a small initial random sample of patients, half with good prognosis or non-tumor tissues and half with poor prognosis or tumor-like tissues, either (5,5) or (10,10). Following application of the sequential algorithm, the misclassification error for the sequentially trained classifier is computed based on the remaining portion of the sample (which serves as hold-out test data) by comparing the predicted and clinical outcomes. For each set of fixed criteria such as $\varepsilon > 0$, $\alpha = 0.05$ and the maximum number of samples allowed in the sequential procedure, we carry out the sequential procedure 1000 times and report the mean number of sequential steps and the mean misclassification error.

5.1 Example 1: LDA classifier on small number of genes for breast-cancer patient prognosis

An LDA classifier is trained with the sequential procedure on the three genes most highly correlated with the prognosis based on the total of 295 samples. We are aware of the selection bias in such a gene selection procedure based on all 295 samples (Ambroise and McLachlan, 2002), but proceed in this way so as to demonstrate our method without incorporating a gene selection procedure. Example 3 will demonstrate the method incorporating a gene selection procedure. We choose a small number of genes to train the classifier for two reasons. First, biologists and clinicians often have only a small number of experimentally identified genes that play a major role, which depends largely on investigator’s subjective judgement and is potentially subject to selection bias as well. Second, the amount of training data increases substantially with the number of classifier features.

Table 3 shows that, as the threshold ε decreases from 0.3 to 0.15, the mean misclassification error decreases and the minimum and mean number of sequential steps increase. In both cases, the maximum number of steps allowed in the sequential procedure is achieved. The increase in the

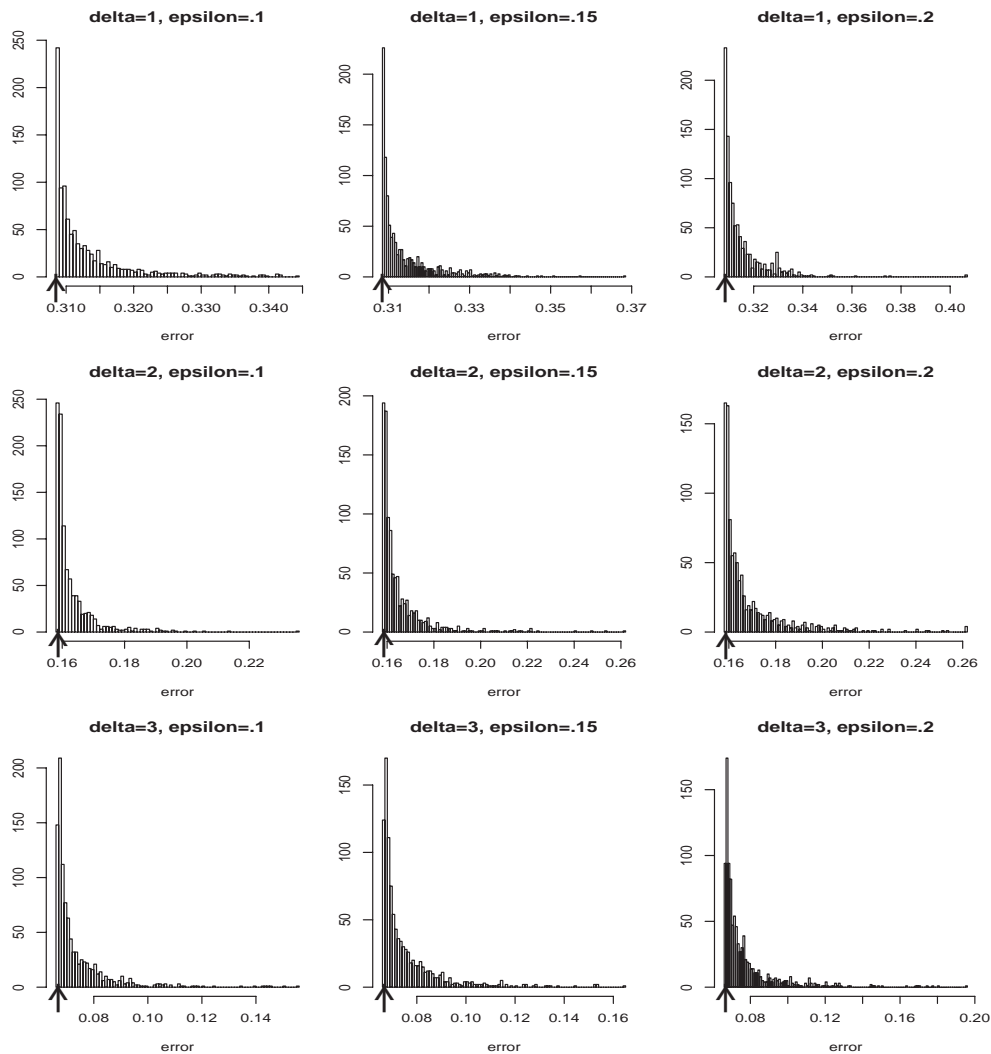


Fig. 1. Histogram of misclassification errors in 1000 runs for given Δ and ε in the simulation study with maximum steps $M = 90$. The arrow in the bottom left corner of each plot indicates the Bayes error between Normal $(0, 1)$ and Normal $(\Delta, 1)$ with specified Δ . Top panels, $\Delta = 1$; Middle panels, $\Delta = 2$; Bottom panels, $\Delta = 3$; Left panels, $\varepsilon = 0.1$; Central panels, $\varepsilon = 0.15$; and Right panels, $\varepsilon = 0.2$.

maximum number of steps from 50 to 80 is due to the increase in the maximum number of steps allowed in the sequential procedure. With the initial sample size increasing from $(5, 5)$ to $(10, 10)$, the mean misclassification error and the mean number of steps decrease. Although the initial sample size does not contribute directly to the stopping rule, a larger initial sample contributes to better classifier training and therefore may lead to early stopping.

5.2 Example 2: K -nearest neighbors (KNN) classifiers on small number of genes for breast-cancer patient prognosis

We fit 3NN and 5NN classifiers separately on the 5 genes most highly correlated with patient prognosis based on the full sample of 295 patients. The procedure is similar to Example 1

and the results are reported in Tables 4 and 5. It is interesting to observe in the tables that although the 3NN and 5NN classification rules have about the same mean number of steps and yield about the same misclassification errors, the 5NN classifier yields a slightly smaller mean number of steps and has slightly smaller errors than the 3NN classifier.

The mean misclassification errors of ~ 0.22 – 0.25 reported in Tables 3–5 by the sequentially trained classifiers are consistent with the errors reported by Braga-Neto and Dougherty (2004).

5.3 Example 3: 3NN classifier on 10 sequentially selected principal components for breast-tumor characterization

We study the 3NN classifier on 10 principal components of the gene expressions for breast-tumor characterization. First,

Table 2. Misclassification error, the SD, minimum, maximum and mean (\bar{N}) numbers of sequential steps and the SD in training LDA classifier by population mean Δ and stopping threshold ε for fixed $\alpha = 0.05$ and maximum sequential steps 40

| Δ | Bayes error ^a | | ε | | | |
|----------|--------------------------|-------------------|-----------------|-----------------|-----------------|-----------------|
| | | | 0.05 | 0.1 | 0.15 | 0.2 |
| 1 | 0.3085 | Err (SD) | 0.3172 (0.0101) | 0.3165 (0.0096) | 0.3171 (0.0100) | 0.3173 (0.0105) |
| | | (min, max) | (40, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (40, 0) | (39, 4) | (37, 9) | (32, 14) |
| 1.3 | 0.2578 | Err (SD) | 0.2651 (0.0098) | 0.2653 (0.0101) | 0.2652 (0.0106) | 0.2676 (0.0126) |
| | | (min, max) | (22, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (40, 1) | (38, 7) | (34, 12) | (29, 15) |
| 1.5 | 0.2266 | Err (SD) | 0.2323 (0.0078) | 0.2332 (0.0095) | 0.2342 (0.0114) | 0.2351 (0.0126) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (40, 1) | (37, 9) | (32, 13) | (26, 15) |
| 2 | 0.1587 | Err (SD) | 0.1636 (0.0069) | 0.1649 (0.0103) | 0.1672 (0.0133) | 0.1683 (0.0144) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (39, 3) | (33, 11) | (24, 14) | (19, 13) |
| 2.3 | 0.1251 | Err (SD) | 0.1297 (0.0076) | 0.1319 (0.0117) | 0.1338 (0.0144) | 0.1355 (0.0176) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (38, 5) | (30, 12) | (21, 13) | (15, 10) |
| 2.5 | 0.1056 | Err (SD) | 0.1101 (0.0062) | 0.1122 (0.0106) | 0.1142 (0.0136) | 0.1158 (0.0156) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (38, 5) | (28, 12) | (20, 12) | (14, 9) |
| 3 | 0.0668 | Err (SD) | 0.0710 (0.0063) | 0.0737 (0.0118) | 0.0761 (0.0178) | 0.0785 (0.0199) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 40) |
| | | (\bar{N} , SD) | (35, 7) | (24, 11) | (16, 8) | (12, 5) |
| 4 | 0.0228 | Err (SD) | 0.0266 (0.0071) | 0.0287 (0.0113) | 0.0305 (0.0153) | 0.0320 (0.0168) |
| | | (min, max) | (21, 40) | (11, 40) | (8, 40) | (6, 28) |
| | | (\bar{N} , SD) | (34, 7) | (23, 7) | (16, 5) | (12, 3) |

^aBayes error between the two populations Normal (0, 1) and Normal (Δ , 1) was calculated with $e_B = 1 - \Phi(\Delta/2)$, where Φ is the cumulative distribution function of the distribution Normal (0, 1).

Table 3. Misclassification error, its SD, minimum, maximum, mean numbers of sequential steps and the SD in training the LDA classifier on the three most highly correlated genes with breast-cancer patient’s prognosis (van de Vijver et al., 2002)

| Initial size | (ε , M) | Mean error (SD) | Mean steps (SD) | Min | Max |
|--------------|-------------------------|-----------------|-----------------|-----|-----|
| (5, 5) | (0.30, 50) | 0.2689 (0.0632) | 17.7 (15.8) | 4 | 50 |
| | (0.20, 50) | 0.2402 (0.0474) | 34.2 (18.9) | 6 | 50 |
| | (0.15, 80) | 0.2235 (0.0379) | 64.3 (27.9) | 8 | 80 |
| (10,10) | (0.30, 50) | 0.2396 (0.0348) | 13.9 (13.9) | 4 | 50 |
| | (0.20, 50) | 0.2290 (0.0285) | 29.4 (19.5) | 6 | 50 |
| | (0.15, 80) | 0.2202 (0.0285) | 58.1 (31.0) | 8 | 80 |

M is the predetermined maximum number of steps allowed.

Table 4. Misclassification error, the SD, minimum, maximum, mean numbers of sequential steps and the SD in training the 3NN classifier on the five most highly correlated genes with breast-cancer patient’s prognosis (van de Vijver et al., 2002)

| Initial size | (ε , M) | Mean error (SD) | Mean steps (SD) | Min | Max |
|--------------|-------------------------|-----------------|-----------------|-----|-----|
| (5, 5) | (0.30, 50) | 0.2519 (0.0443) | 16.8 (16.0) | 4 | 50 |
| | (0.20, 50) | 0.2443 (0.0348) | 31.6 (19.5) | 6 | 50 |
| | (0.15, 80) | 0.2373 (0.0316) | 60.7 (30.1) | 8 | 80 |
| (10,10) | (0.30, 50) | 0.2500 (0.0379) | 16.5 (15.6) | 4 | 50 |
| | (0.20, 50) | 0.2433 (0.0348) | 31.6 (19.6) | 6 | 50 |
| | (0.15, 80) | 0.2358 (0.0285) | 58.3 (31.3) | 8 | 80 |

M is the maximum number of steps allowed.

we remove the genes with missing expression levels and keep 4982 genes with no missing values. Second, we order the 4982 genes by the correlation between the tumor class and each individual gene expression based on the entire 84 samples, and choose 445 genes that have a correlation coefficient of absolute value greater than 0.4. These 445 genes are used to train the 3NN classifier in the sequential procedure. We

also incorporate feature extraction in this procedure. At each sequential step, we compute the 10 largest principal components based on the current available sample of gene expressions and train a 3NN classifier based on the principal components. Table 6 reports the mean misclassification error, the mean, minimum and maximum number of sequential steps over 1000 runs for each fixed pair of (ε , M), where M is the maximum

Table 5. Misclassification error, the SD, minimum, maximum, mean numbers of sequential steps and the SD in training the 5NN classifier on the five most highly correlated genes with breast-cancer patient's prognosis (van de Vijver *et al.*, 2002)

| Initial size | (ε, M) | Mean error (SD) | Mean steps (SD) | Min | Max |
|--------------|--------------------|-----------------|-----------------|-----|-----|
| (5, 5) | (0.30, 50) | 0.2479 (0.0411) | 15.2 (14.6) | 4 | 50 |
| | (0.20, 50) | 0.2388 (0.0348) | 30.9 (19.6) | 6 | 50 |
| | (0.15, 80) | 0.2303 (0.0285) | 59.3 (30.7) | 8 | 80 |
| (10,10) | (0.30, 50) | 0.2413 (0.0348) | 14.6 (14.6) | 4 | 50 |
| | (0.20, 50) | 0.2359 (0.0316) | 29.7 (19.6) | 6 | 50 |
| | (0.15, 80) | 0.2288 (0.0285) | 57.8 (31.1) | 8 | 80 |

M is the maximum number of steps allowed.

Table 6. Misclassification error, the SD, minimum, maximum, mean numbers of sequential steps and the SD in training the 3NN classifier based on sequentially selected 10 largest principal components of expressions of the 445 most highly correlated genes with the tissue class in the breast-tumor characterization study (Perou *et al.*, 2000)

| Initial size | (ε, M) | Mean error (SD) | Mean steps (SD) | Min | Max |
|--------------|--------------------|-----------------|-----------------|-----|-----|
| (5, 5) | (0.20, 30) | 0.1038 (0.0538) | 12.8 (6.6) | 6 | 30 |
| | (0.15, 30) | 0.0981 (0.0538) | 17.0 (8.1) | 8 | 30 |
| | (0.15, 50) | 0.0957 (0.0569) | 19.6 (12.3) | 8 | 50 |
| | (0.10, 50) | 0.0786 (0.0538) | 31.1 (15.7) | 11 | 50 |
| (10,10) | (0.20, 30) | 0.0755 (0.0411) | 11.9 (4.8) | 6 | 30 |
| | (0.15, 30) | 0.0695 (0.0379) | 15.5 (6.6) | 8 | 30 |
| | (0.15, 50) | 0.0677 (0.0379) | 16.5 (8.2) | 4 | 50 |
| | (0.10, 50) | 0.0613 (0.0379) | 25.7 (13.5) | 11 | 50 |

M is the maximum number of steps allowed.

number of sequential steps allowed. It is shown that the misclassification error decreases with the decrease in the threshold ε and increase in the maximum number of steps allowed. By specifying a small threshold ε , a decrease in the misclassification error may be achieved, but at the cost of running more sequential steps. With larger initial samples, (10,10) versus (5,5), the sequential procedure runs a fewer number of steps to stop and yields smaller error.

We have observed different levels of the misclassification errors in the two microarray studies, $\sim 25\%$ in the breast-cancer prognosis study and $< 10\%$ in the breast-tumor characterization study. This reflects the fact, as we mentioned earlier in Section 2, that the misclassification error depends on many factors, including Bayes error, the number of predictors in the model, and most importantly, the scientific context of the problem.

6 CONCLUSION

We have studied a sequential classification procedure and provided a stopping rule. The procedure recruits subjects sequentially, updates the classification rule at each step and stops with certain predetermined confidence level $(1 - \alpha)$

such that a new subject will have a probability less than a small threshold $\varepsilon > 0$ to be misclassified by the classification rule. Simulation studies show that the procedure usually achieves its goal of stopping prior to reaching the maximum allowable sample size, and the gain is over a fairly wide range of ε . We have applied the procedure to two microarray datasets. In particular, we have applied it in the context of feature extraction. It has been shown to yield error rates close to those achievable for fixed sample sizes.

This procedure has several advantages over classical sample size calculations: (1) it updates the classification rule sequentially and thus depends on the study subjects rather than relying on distributions of primary measurements from other studies that may differ greatly, as do the classical sample size calculations; (2) it assesses the stopping criteria at each sequential step and thus can substantially reduce cost via early stopping; (3) it ensures the required sample size to achieve significance, whereas classical sample size calculations may lead to smaller than required sample sizes and therefore miss the expected significance due to inaccurate estimation; and (4) it is not restricted to any particular classification rule and therefore applies to any parametric and non-parametric method, such as LDA, KNN, classification and regression trees (CART), etc.

Although we have considered the classification of binary outcomes in this study, our procedure also applies to multiple class outcomes as well, since one can still define the misclassification indicator Q_i and its conditional probability π_i with multiple classes, and these constitute the only information used to derive the stopping rule.

ACKNOWLEDGEMENTS

W.J.F. was supported by a grant from the National Cancer Institute (CA-90301). E.R.D. was supported by the Translational Genomics Research Institute. R.J.C. and B.M. were supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

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., USA*, **99**, 6562–6566.
- Braga-Neto, U.M. and Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.
- Emerson, S.S. and Fleming, T.R. (1990) Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875–892.
- Hall, P. and Heyde, C.C. (1980) *Martingale Limit Theory and Its Application*. Academic Press Inc., New York.
- Knight, K. (2000) *Mathematical Statistics*. Chapman and Hall/CRC, New York.

Lai, T.L. (1997) On optimal stopping problems in sequential hypothesis testing. *Statist. Sinica*, **7**, 33–51.

Liu, A. and Hall, W.J. (1999) Unbiased estimation following a group sequential test. *Biometrika*, **86**, 71–78.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H. Akslen, L.A. et al. (2000) Molecular portraits of Human Breast Tumours. *Nature*, **406**, 747–752.

Pinheiro, J.C. and DeMets, D.L. (1997) Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika*, **84**, 831–845.

Shorack, G.R. (2000) *Probability for Statisticians*. Springer, New York.

Todd, S. and Whitehead, J. (1996) Point and interval estimation following a sequential clinical trial. *Biometrika*, **83**, 453–461.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J. et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Whitehead, J. (1986) On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, **73**, 573–581.

A APPENDIX

A.1 Derivation of the stopping rule

Let $X_i = Q_i - \pi_i$, and \mathcal{F}_n be the σ -field generated by Y_n . Then $(\sum_{i=1}^n X_i, \mathcal{F}_n)$ is a zero-mean martingale (Hall and Heyde, 1980). Let $s_n^2 = \text{var}(\sum_{i=1}^n X_i)$. By the variance of conditional expectation, one is ready to conclude the convergence in probability

$$n^{-1} s_n^2 = n^{-1} \sum_{i=1}^n E[\pi_i(1 - \pi_i)] \rightarrow_p \pi_\infty(1 - \pi_\infty).$$

It is thus ready to prove with Chebyshev's Inequality (Knight, 2000, p. 123) that $s_n^{-1} \sum_{i=1}^n X_i^2 \rightarrow_p 1$ and $E(\max_{1 \leq i \leq n} s_n^{-1} |X_i|) \rightarrow 0$. By the martingale central limit theorem (CLT) (Shorack, 2000, pp. 529–530), one has the convergence in distribution

$$s_n^{-1} \sum_{i=1}^n X_i \rightarrow_d N(0, 1).$$

Let $\kappa_n = n^{-1} \sum_{i=1}^n \pi_i(1 - \pi_i)$. By the CLT, $n^{1/2}(\bar{Q}_n - \bar{\pi}_n)/\kappa_n \rightarrow_d N(0, 1)$ in distribution. Thus with probability approaching $1 - \alpha$,

$$\bar{\pi}_n \leq \bar{Q}_n + z_{1-\alpha} \kappa_n / n^{-1/2}. \quad (\text{A.1})$$

For the stopping rule, we want to find a sample size N such that the probability $\text{Prob}(\pi_N \leq \varepsilon) \geq 1 - \alpha$.

For large m , by the weak monotonicity of π_n ,

$$\begin{aligned} \frac{\pi_1 + \cdots + \pi_m}{m} &= \frac{\pi_1 + \cdots + \pi_{m-p}}{m} + \frac{\pi_{m-p+1} + \cdots + \pi_m}{m} \\ &\geq \pi_m + C_0, \end{aligned}$$

where $C_0 = m^{-1} \sum_{i=1}^p [\pi_{m-p+i} - \pi_m]$ converges to 0. Hence for large m , $\bar{\pi}_m = m^{-1} [\pi_1 + \cdots + \pi_m] \geq \pi_m$ in probability, i.e. $\text{P}(\bar{\pi}_m \geq \pi_m) \rightarrow 1$. Particularly, $\text{P}(\bar{\pi}_N \geq \pi_N) \rightarrow 1$. Hence by (A.1)

$$\text{P}\left(\pi_N \leq \bar{Q}_N + z_{1-\alpha} \kappa_N / N^{-1/2}\right) \rightarrow 1 - \alpha. \quad (\text{A.2})$$

We now show that $\text{P}(\hat{\kappa}_N \geq \kappa_N) \rightarrow 1$ as $N \rightarrow \infty$. We first show that $\hat{\pi}_n \geq \pi_n$ in probability for large n . Since $s_n / \sqrt{n} \rightarrow \sqrt{\pi_\infty(1 - \pi_\infty)}$ and the convergence in distribution

$$\frac{n}{s_n} (\bar{Q}_n - \bar{\pi}_n) \rightarrow_d N(0, 1),$$

it implies that $\bar{Q}_n = \bar{\pi}_n + o(n^{-1/2})$ for large n . Thus $\hat{\pi}_n = \bar{\pi}_n + o(n^{-1/2}) \geq \pi_n$ in probability. Without loss of generality, we assume $\hat{\pi}_n \leq 1/2$ a.s. Note that if $0 < a \leq b \leq 1/2$, then $a(1 - a) \leq b(1 - b)$. Thus $\hat{\pi}_n(1 - \hat{\pi}_n) \geq \pi_n(1 - \pi_n)$ in probability for large n .

Let M_0 be a large integer such that $M_0 < N$ and $\hat{\pi}_i \geq \pi_i$ in probability for $i \geq M_0$.

$$\begin{aligned} \hat{\kappa}_N - \kappa_N &= \frac{\left(\sum_{i=1}^{M_0} + \sum_{i=M_0+1}^N\right) \hat{\pi}_i(1 - \hat{\pi}_i)}{N} \\ &\quad - \frac{\left(\sum_{i=1}^{M_0} + \sum_{i=M_0+1}^N\right) \pi_i(1 - \pi_i)}{N} \\ &\geq \frac{\sum_{i=1}^{M_0} \{\hat{\pi}_i(1 - \hat{\pi}_i) - \pi_i(1 - \pi_i)\}}{N} = \frac{C_1}{N} \end{aligned}$$

holds in probability, where $C_1 = \sum_{i=1}^{M_0} [\hat{\pi}_i(1 - \hat{\pi}_i) - \pi_i(1 - \pi_i)]$ is bounded by a finite number M_0 . Hence $\text{P}(\hat{\kappa}_N \geq \kappa_N) \rightarrow 1$ and by (A.2)

$$\text{P}\left(\pi_N \leq N^{-1} \sum_{i=1}^N Q_i + z_{1-\alpha} \hat{\kappa}_N / N^{1/2}\right) \rightarrow 1 - \alpha$$

holds. This completes the proof for the theorem.

Notice that the stopping rule requires $\hat{\kappa}_n > 0$. However, a series of consecutive perfect classifications from the beginning of the sequential procedure yields $\hat{\kappa}_n = 0$. Meanwhile, a long sequence of perfect classifications indicates good performance of classifiers and thus should invoke stopping. This is amended in the stopping rule by setting the number of consecutive perfect classifications $N_0 = \log(\alpha) / \log(1 - \varepsilon)$ with a conservative assumption that the perfect classifications are independent with least probability $(1 - \varepsilon)$ for each.