BALANCED BOOTSTRAP ESTIMATORS FOR THE PROBABILITY OF MISCLASSIFICATIONS IN DISCRIMINANT ANALYSIS

I WAYAN MANGKU

Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University
Jl. Meranti, Kampus IPB Darmaga, Bogor, 16680 Indonesia

Abstract. In this paper we propose new error rate estimators based on balanced bootstrap technique, which are expected to perform better than the existing estimators. These estimators can be computed by means of separate or mixture resampling methodology. We consider both of them.

Key words: Discriminant analysis, classification rule, probability of misclassification, actual error rate, balanced bootstrap estimator.

1. Introduction

The major problem in estimating the error rate in discriminant analysis arises when there are no data available beyond those used to define and estimate the classification rule. The resubstitution method has been reported optimistically biased because of using the same data to construct as well as to evaluate the classification rule.

One other alternative is using parametric estimators. When the parent populations are multivariate normal, on average, the OS, L, and M estimators are the best for estimating the actual error rate. However, the performance of these estimators deteriorate when the parent populations are not normal. So, when the normality assumption is questioned, we still need a better estimator.

Another alternative is using the empirical estimators such as the ones based on cross-validation, jackknife, and bootstrap. The best estimators among these empirical techniques (see Efron (1983), Snapinn and Knoke (1985), Ganeshanandam and Krzanowski (1990) and Mangku (1992)) are the U, $\bar{U}$, JK, and 0.632 estimators. These U, $\bar{U}$, and JK estimators are basically based on the leave-one-out technique. Since this procedure holds out one observation at a time, in turn, until each observation has been held once, the maximum number of pseudo data created here is the same as the original sample size. Because of this

fact, the performance of these estimators deteriorate when the sample sizes become small. In other words, when the sample sizes are small, we still need a better estimator.

In the case of small samples, we expect the bootstrap based technique (0.632 estimator) to behave better, since the number of pseudo data that can be generated here is almost independent of the sample sizes. The number of bootstrap samples that can be re-sampled (with replacement) from a sample of size n is n^n . Here, we can notice that the number of pseudo data sets, namely n^n, is much larger than the size of the original sample, n, even for small values of n. However, as reported by Ganeshanandam and Krzanowski (1990), this 0.632 estimator perform poorly when the two populations are not closer together.

So we still need a better error rate estimator, particularly when the populations are not normal, when the populations are not closer together, and when the sample sizes are small.

In this paper we propose new error rate estimators based on balanced bootstrap technique, which are expected to perform better than the existing estimators. These estimators can be computed by means of both separate and mixture resampling methodology. We shall consider both of these cases. Before discussing the proposed estimators, the basic idea of the balanced bootstrap technique is reviewed in section 2. Derivation and algorithm of the balanced bootstrap estimator using separate bootstrap samplings are presented in section 3, while section 4 gives explanations of the balanced bootstrap estimator using mixture bootstrap samplings. Some further discussions are given in section 5.

2. The Balanced Bootstrap Technique

Some improvements and modifications of the ordinary bootstrap methodology have been reported in the literature. Efron (1983) introduced randomized bootstrap and double bootstrap to correct the bias of the ordinary bootstrap. The randomized bootstrap is a simple modification of the ordinary bootstrap, appropriate when the data are dichotomous. Suppose we have n observations $x_1, x_2, \ldots, x_n$, then we put mass $1/n$ at each x_i $(i = 1, 2, \ldots, n)$ as well as at each of the complementary point $(1 - x_i)$ to construct the empirical probability distribution $\hat{F}$. The bootstrapping procedure is then performed according to this empirical probability distribution. While in the double bootstrap, the procedure involve two layers of bootstrapping. Here, the second layer of bootstrap sample is obtained by resampling the first layer of bootstrap sample. This second layer of bootstrap sample is then used to correct the bias of the ordinary bootstrap.

Later, some other modifications such as balanced resampling, importance resampling, antithetic resampling, and nested resampling for the bootstrap also have been introduced (see Davison, Hinkley, and Schechtman (1986), Gleason (1988), Johns (1988), Hall (1989a,..., 1990),

Hall et al. (1989), Hinkley and Shi (1989), Efron (1990), and Graham et al. (1990)). The most popular and widely applicable modification among these is due to the balanced resampling. This is a technique where each sample observation occurs with the same frequency in all of the bootstrap samples.

The bootstrap sample obtained from balanced resampling has computational advantages over the ordinary one. Also, in the estimation procedures the balanced bootstrap resampling reduces the simulation error by controlling a linear approximation to the estimate of the parameter (Davison, Hinkley, and Schechtman (1986), Gleason (1988)). Furthermore, Hall (1989a) showed that the balanced bootstrap algorithm has mean squared error of $(bn^2)^{-1}$ which represents a significant improvement over the ordinary one, which has mean squared error of only $(bn)^{-1}$. Here n is the sample size and b is the number of bootstrap samples.

The need for balanced bootstrap simulation can be illustrated as follows. Consider the simple case of bootstrapping the average

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

of random variables $Y_1, Y_2, \ldots, Y_n$. Let $y_1, y_2, \ldots, y_n$ be the observed values of $Y_1, Y_2, \ldots, Y_n$. Suppose that we estimate the bias and the variance of $\bar{Y}$ by ordinary bootstrap simulation; these estimates are given respectively by

$$Bias^* \;=\; \bar{y}^*_{(b)} - \bar{y} \tag{2.1}$$

and

$$s^{*2}_{(b)} \;=\; \frac{1}{b-1}\sum_{k=1}^{b}\left(\bar{y}^*_k - \bar{y}^*_{(b)}\right)^2. \tag{2.2}$$

Here $\bar{y}^*_k$'s $(k = 1, 2, \ldots, b)$ are the averages from the bootstrap samples and $\bar{y}^*_{(b)}$ is the average of these b values and $\bar{y}$ is the average of $y_1, y_2, \ldots, y_n$. Evidently the nonzero bias estimate "$Bias^*$" will be misleading because the true bias of $\bar{Y}$ is zero. Removing this error requires only that each datum y_j $(j = 1, 2, \ldots, n)$ occurs equally often in the aggregate of all b bootstrap samples. This simple balance also has concomitant effect of slightly reducing the probable error in $s^{*2}_{(b)}$ (Davison, Hinkley, and Schechtman, 1986).

The general procedure to perform balanced bootstrap resampling can be summarized as follows. Suppose that we have a data set $Y = \{y_1, y_2, \ldots, y_n\}$ of size n, and we wish to generate b balanced bootstrap samples. Theoretically, the procedure is

(a) Make b copies of each datum $y_1, y_2, \ldots, y_n$, so we will have a new big data set of size bn.

(b) Draw a random sample of size n from the new big data set without replacement. This is the first balanced bootstrap sample.
(c) Repeat step (b) until we have b balanced bootstrap samples.

(Graham, et al., 1990).

3. Estimators based on separate samplings

This section explains the balanced bootstrap procedure for estimating the misclassification error rate in discriminant analysis using separate bootstrap samplings approach. Suppose we wish to generate b balanced bootstrap samples, and the algorithm can be described as follows:

(a) Let us first consider the generation of bootstrap samples from the training set $\underline{\mathbf{t}}_1$. Construct a list L of length bn_1 (i.e. b times n_1) by concatenating $l_1, l_2, \ldots, l_b$, where each l_k $(k = 1, 2, \ldots, b)$ is a set of indices of the form $\{1, 2, \ldots, n_1\}$. Then, randomly permute L to produce a new list L', and successively divide L' into b new sets of indices $l'_1, l'_2, \ldots, l'_b$ each of size n_1. Note here that each of these l'_k's may contain more than one replication of some numbers from $\{1, 2, \ldots, n_1\}$, hence leaving no representation of some other numbers of this set in them. The balanced bootstrap samples are then given by $\underline{\mathbf{t}}^*_{11}, \underline{\mathbf{t}}^*_{12}, \ldots, \underline{\mathbf{t}}^*_{1b}$, where the elements of $\underline{\mathbf{t}}^*_{1k}$ are the elements of $\underline{\mathbf{t}}_1$ corresponding to the indices of l'_k $(k = 1, 2, \ldots, b)$. Repeat the above process with n_1 and $\underline{\mathbf{t}}_1$ replaced respectively by n_2 and $\underline{\mathbf{t}}_2$, to obtain the balanced bootstrap samples $\underline{\mathbf{t}}^*_{11}, \underline{\mathbf{t}}^*_{12}, \ldots, \underline{\mathbf{t}}^*_{1b}$ from training set $\underline{\mathbf{t}}_2$.
(b) Based on the bootstrap training sets $\underline{\mathbf{t}}^*_k = \{\underline{\mathbf{t}}^*_{1k}, \underline{\mathbf{t}}^*_{2k}\}$, construct classification rules $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*_k)$, for $k = 1, 2, \ldots, b$.
(c) Classify the individuals in the original training sample $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ which are not drawn into the k-th balanced bootstrap sample $\underline{\mathbf{t}}^*_k = \{\underline{\mathbf{t}}^*_{1k}, \underline{\mathbf{t}}^*_{2k}\}$ using the k-th classification rule $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*_k)$. Let β_k be the number of individuals in $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ which are not drawn into $\underline{\mathbf{t}}^*_k = \{\underline{\mathbf{t}}^*_{1k}, \underline{\mathbf{t}}^*_{2k}\}$. Let also α_k be the number of individuals in $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ which are not drawn into $\underline{\mathbf{t}}^*_k = \{\underline{\mathbf{t}}^*_{1k}, \underline{\mathbf{t}}^*_{2k}\}$ and are misclassified by $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}^*_k)$. Then compute

$$\xi = \left(\sum_{k=1}^{b} \alpha_k\right) / \left(\sum_{k=1}^{b} \beta_k\right).$$

The balanced bootstrap estimator using separate bootstrap samplings is then given by

$$\hat{P}(SBB) = (1 - C_s)\hat{P}(R) + C_s\xi, \tag{3.1}$$

with SBB means *Separate Balanced Bootstrap*. Here, $\hat{P}(R)$ is the *overall resubstitution error rate* (see Smith (1947)), and C_s represents the probability that the observation $\underline{\mathbf{x}}_{ij}$ in the original training sample $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ is selected into the balanced bootstrap

sample $\underline{\mathbf{t}}_k^* = \{\underline{\mathbf{t}}_{1k}^*, \underline{\mathbf{t}}_{2k}^*\}$, using separate resampling methodology (for $i = 1, 2$; $j = 1, 2, \ldots, n_i$; and $k = 1, 2, \ldots, b$).

The use of the coefficient C_s in equation (3.1) follows Efron's (1983) assumption about the distribution of the distance between the point at which the classification rule is applied and the nearest point in the training samples. Efron (1983) assumes that this probability distribution is equal to the probability that the point at which the rule is applied is included in the bootstrap samples. For the case of balanced bootstrap with separate bootstrap samplings, the expression of this probability can be derived as below.

Recall the procedure to resample $\underline{\mathbf{t}}_1$ to obtain the balanced bootstrap samples $\underline{\mathbf{t}}_{1k}^*$, $(k = 1, 2, \ldots, b)$ as derived in step (a) above. From this process, it is clear that the probability that an individual $\underline{\mathbf{x}}_{1j}$ in $\underline{\mathbf{t}}_1$ is included in the balanced bootstrap sample $\underline{\mathbf{t}}_{1k}^*$ is the same as the probability that an element z in l is selected into l'_k, z being any number from $\{1, 2, \ldots, n_1\}$. For a particular value of z, L' will consists of b duplicates of z and $(bn_1 - b)$ other numbers. Since we divide L' into b l'_k's $(k = 1, 2, \ldots, b)$, each of size n_1, the probability that z is not included in l'_k, follows a hypergeometric model, and is given by

$$F(n_1, b) = \frac{\begin{pmatrix} b \\ 0 \end{pmatrix} \begin{pmatrix} n_1 b - b \\ n_1 \end{pmatrix}}{\begin{pmatrix} n_1 b \\ n_1 \end{pmatrix}}, \tag{3.2}$$

where

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a!}{b!(a-b)!}.$$

Hence, the probability that the individual $\underline{\mathbf{x}}_{1j}$ in $\underline{\mathbf{t}}_1$ is not included in the balanced bootstrap sample $\underline{\mathbf{t}}_{1k}^*$ is given by equation (3.2), for any $j = 1, 2, \ldots, n_1$, and $k = 1, 2, \ldots, b$. Using similar explanations as above, we can deduce that the probability for any individual $\underline{\mathbf{x}}_{2j}$ in $\underline{\mathbf{t}}_2$ is not included in the balanced bootstrap sample $\underline{\mathbf{t}}_{2k}^*$, for any $j = 1, 2, \ldots, n_2$ and $k = 1, 2, \ldots, b$, is given by

$$F(n_2, b) = \frac{\begin{pmatrix} b \\ 0 \end{pmatrix} \begin{pmatrix} n_2 b - b \\ n_2 \end{pmatrix}}{\begin{pmatrix} n_2 b \\ n_2 \end{pmatrix}}. \tag{3.3}$$

Note that, since $\underline{\mathbf{t}}_{1k}^*$'s are resampled from $\underline{\mathbf{t}}_1$ only, the probability that the individual $\underline{\mathbf{x}}_{2j}$ in $\underline{\mathbf{t}}_2$ is not included in the balanced bootstrap sample $\underline{\mathbf{t}}_{1k}^*$ is equal to 1, for $j = 1, 2, \ldots, n_2$ and $k = 1, 2, \ldots, b$. It is also obvious that the probability that the individual $\underline{\mathbf{x}}_{1j}$ in $\underline{\mathbf{t}}_1$ is not included in the balanced bootstrap sample $\underline{\mathbf{t}}_{2k}^*$ is equal to 1, since $\underline{\mathbf{t}}_{2k}^*$'s are also resampled from $\underline{\mathbf{t}}_2$ only, for $j = 1, 2, \ldots, n_1$ and $k = 1, 2, \ldots, b$.

Since $\underline{\mathbf{t}}_k^* = \{\underline{\mathbf{t}}_{1k}^*, \underline{\mathbf{t}}_{2k}^*\}$, the probability for an individual $\underline{\mathbf{x}}_{1j}$ is not included in the k-th balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ ($j = 1, 2, \ldots, n_1$; $k = 1, 2, \ldots, b$) is given by

$$\begin{aligned}\mathsf{P}\left(\underline{\mathbf{x}}_{1j} \notin \underline{\mathbf{t}}_k^*\right) &= \mathsf{P}\left(\underline{\mathbf{x}}_{1j} \notin \underline{\mathbf{t}}_{1k}^*\right)\ \mathsf{P}\left(\underline{\mathbf{x}}_{1j} \notin \underline{\mathbf{t}}_{2k}^*\right) = F\left(n_1, b\right)(1) = F\left(n_1, b\right) \\ &= \frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_1 b - b \\ n_1 \end{array}\right)}{\left(\begin{array}{c} n_1 b \\ n_1 \end{array}\right)}. \end{aligned} \tag{3.4}$$

Using the same argument we also find that the probability that the individual $\underline{\mathbf{x}}_{2j}$ is not included in the k-th balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ ($j = 1, 2, \ldots, n_2$; $k = 1, 2, \ldots, b$) is given by

$$\begin{aligned}\mathsf{P}\left(\underline{\mathbf{x}}_{2j} \notin \underline{\mathbf{t}}_k^*\right) &= \mathsf{P}\left(\underline{\mathbf{x}}_{2j} \notin \underline{\mathbf{t}}_{1k}^*\right)\ \mathsf{P}\left(\underline{\mathbf{x}}_{2j} \notin \underline{\mathbf{t}}_{2k}^*\right) = (1)F\left(n_2, b\right) = F\left(n_2, b\right) \\ &= \frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_2 b - b \\ n_2 \end{array}\right)}{\left(\begin{array}{c} n_2 b \\ n_2 \end{array}\right)}. \end{aligned} \tag{3.5}$$

Since the entire set $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ is a mixture of n_1 individuals belong to $\underline{\mathbf{t}}_1$ and n_2 individuals belong to $\underline{\mathbf{t}}_2$, for any individual $\underline{\mathbf{x}}_{ij} \in \underline{\mathbf{t}}$ ($i = 1, 2; j = 1, 2, \ldots, n_i$), the probability that this individual is not included in the k-th balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ (k=1,2,...,b) is given by

$$\begin{aligned}\bar{F}\left(n_1, n_2, b\right) &= \frac{n_1}{n_1 + n_2} F\left(n_1, b\right) + \frac{n_2}{n_1 + n_2} F\left(n_2, b\right) \\ = \frac{n_1}{n_1 + n_2} &\frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_1 b - b \\ n_1 \end{array}\right)}{\left(\begin{array}{c} n_1 b \\ n_1 \end{array}\right)} + \frac{n_2}{n_1 + n_2} \frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_2 b - b \\ n_2 \end{array}\right)}{\left(\begin{array}{c} n_2 b \\ n_2 \end{array}\right)}. \end{aligned} \tag{3.6}$$

Finally, we find the probability that any individual $\underline{\mathbf{x}}_{ij}$ in $\underline{\mathbf{t}}$ ($i = 1, 2; j = 1, 2, \ldots, n$) is included in the k-th balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ ($k = 1, 2, \ldots, b$) as

$$\begin{aligned}\theta\left(n_1, n_2, b\right) &= 1 - \bar{F}\left(n_1, n_2, b\right) \\ &= 1 - \frac{n_1}{n_1 + n_2} \frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_1 b - b \\ n_1 \end{array}\right)}{\left(\begin{array}{c} n_1 b \\ n_1 \end{array}\right)} \\ &- \frac{n_2}{n_1 + n_2} \frac{\left(\begin{array}{c} b \\ 0 \end{array}\right)\left(\begin{array}{c} n_2 b - b \\ n_2 \end{array}\right)}{\left(\begin{array}{c} n_2 b \\ n_2 \end{array}\right)}. \end{aligned} \tag{3.7}$$

Having established the probability expression in (3.7), we can now define the finite version of the balanced bootstrap estimator using separate bootstrap samplings (by replacing C_s in equation (3.1) with $\theta(n_1, n_2, b)$ as,

$$\hat{P}(FSB) = (1 - \theta\,(n_1, n_2, b))\,\hat{P}(R) + \theta\,(n_1, n_2, b)\,\xi, \tag{3.8}$$

with FSB means *Finite Separate Balanced*.

To define the infinite version of the balanced bootstrap estimator with separate sampling procedure, we have to find the convergence value of $\theta(n_1, n_2, b)$ in (3.7) for large values of n_1, n_2, and b. This is given in the following theorem.

Lemma 3.1: For the function $F(n, b)$ given in (3.2) or (3.3), we have

$$\lim_{n\to\infty, b\to\infty} F\,(n, b) = 0.368. \tag{3.9}$$

Proof: By (3.2) or (3.3), we obtain

$$\begin{aligned}
&\lim_{n\to\infty} F\,(n, b) \\
= &\lim_{n\to\infty} \frac{\binom{b}{0}\binom{nb-b}{n}}{\binom{nb}{n}} = \lim_{n\to\infty} \frac{(nb-b)!(nb-n)!}{(nb)!(nb-n-b)!} \\
= &\lim_{n\to\infty} \frac{(nb-b)!(nb-n)(nb-n-1)\ldots(nb-n-b+1)(nb-n-b)!}{(nb)(nb-1)\ldots(nb-b+1)(nb-b)!(nb-n-b)!} \\
= &\lim_{n\to\infty} \frac{(nb-n)(nb-n-1)\ldots(nb-n-b+1)}{(nb)(nb-1)\ldots(nb-b+1)} \\
= &\lim_{n\to\infty} \left(\frac{(nb-n)}{(nb)}\right)\left(\frac{(nb-n-1)}{(nb-1)}\right)\cdots\left(\frac{(nb-n-b+1)}{(nb-b+1)}\right) \\
= &\lim_{n\to\infty} \left(1-\frac{1}{b}\right)\left(1-\frac{n}{(nb-1)}\right)\cdots\left(1-\frac{n}{(nb-b+1)}\right) \\
= &\lim_{n\to\infty} \left(1-\frac{1}{b}\right)\left(1-\frac{1}{b-1/n}\right)\cdots\left(1-\frac{1}{b-(b-1)/n}\right) \\
= &\left(1-\frac{1}{b}\right)\left(1-\frac{1}{b}\right)\cdots\left(1-\frac{1}{b}\right) = \left(1-\frac{1}{b}\right)^b.
\end{aligned}$$

Note that

$$\lim_{b\to\infty}\left(1+\frac{a}{b}\right)^b = e^a$$

where e is the natural logarithm or $e = 2.71828.....$ Hence, for the case $a = -1$, we have

$$\lim_{b\to\infty}\left(1-\frac{1}{b}\right)^{b} = e^{-1} = 0.368.$$

Using these results we find that,

$$\lim_{n\to\infty,b\to\infty} F(n,b) = \lim_{n\to\infty,b\to\infty} \frac{\begin{pmatrix} b \\ 0 \end{pmatrix}\begin{pmatrix} nb-b \\ n \end{pmatrix}}{\begin{pmatrix} nb \\ n \end{pmatrix}} = 0.368. \quad (3.10)$$

Theorem 3.2: For the function $\theta(n_1, n_2, b)$ in (3.7) we have

$$\lim_{n_1\to\infty,n_2\to\infty,b\to\infty} \theta(n_1,n_2,b) = 0.632. \quad (3.11)$$

Proof: Lemma 3.1 implies that,

$$\lim_{n_1\to\infty,b\to\infty} F(n_1,b) = 0.368 \text{ and } \lim_{n_2\to\infty,b\to\infty} F(n_2,b) = 0.368.$$

Hence, for large values of n_1, n_2, and b, the equation (3.6) becomes

$$\begin{aligned} \bar{F}(n_1,n_2,b) &= \frac{n_1}{n_1+n_2}0.368 + \frac{n_2}{n_1+n_2}0.368 \\ &= \left(\frac{n_1}{n_1+n_2}+\frac{n_2}{n_1+n_2}\right)0.368 = 0.368. \end{aligned} \quad (3.12)$$

Finally, for large values of n_1, n_2, and b, the value of function $\theta(n_1, n_2, b)$ in equation (3.7) tends to $(1 - 0.368) = 0.632$. Hence, we may define the infinite version of the balanced bootstrap estimator with separate samplings (by replacing C_s in equation (3.1) with 0.632), as

$$\hat{P}(ISB) = 0.368\hat{P}(R) + 0.632\xi, \quad (3.13)$$

with ISB means *Infinite Separate Balanced*.

4. Estimators based on mixture samplings

This section explains the procedure for estimating the error rate using balanced bootstrap technique with mixture samplings methodology. Suppose that we wish to generate b balanced bootstrap samples, and the algorithm can be summarized as follows:

(a) Construct a list L of length bn by concatenating $l_1, l_2, \ldots, l_b$, where each l_k $(k = 1, 2, \ldots, b)$ is a set of indices from 1 to n, with $n = n_1 + n_2$. Here each l_k has the form

$$l_k = \{1, 2, \ldots, n_1, n_1+1, n_1+2, \ldots, n\},$$

where the first n_1 indices correspond to each element of the first training sample $\underline{\mathbf{t}}_1$, and the remainder correspond to the elements of the second training sample $\underline{\mathbf{t}}_2$.

(b) Randomly permute L to produce a new list L', and successively divide L' into b new sets of indices $l'_1, l'_2, \ldots, l'_b$ each of size n.

(c) The generated balanced bootstrap samples are $\underline{\mathbf{t}}_1^*, \underline{\mathbf{t}}_2^*, \ldots, \underline{\mathbf{t}}_b^*$, where the elements of $\underline{\mathbf{t}}_k^*$ are the elements of $\underline{\mathbf{t}}$ corresponding to the indices of l_k' $(k = 1, 2, \ldots, b)$. Let $\underline{\mathbf{t}}_k^* = \{\underline{\mathbf{t}}_{1k}^*, \underline{\mathbf{t}}_{2k}^*\}$, where $\underline{\mathbf{t}}_{1k}^*$ and $\underline{\mathbf{t}}_{2k}^*$ consist of n_1^* and n_2^* elements respectively from $\underline{\mathbf{t}}_1$ and $\underline{\mathbf{t}}_2$. Here n_i^* is not necessarily equal to n_i, for $i = 1, 2$, but $n_1^* + n_2^* = n_1 + n_2 = n$.

(d) Based on the bootstrap training sets $\underline{\mathbf{t}}_k^*$, construct the classification rules $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}_k^*)$, for $k = 1, 2, \ldots, b$.

(e) Classify the individuals in the original training sample $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ which are not drawn into the k-th balanced bootstrap sample $\underline{\mathbf{t}}_k^* = \{\underline{\mathbf{t}}_{1k}^*, \underline{\mathbf{t}}_{2k}^*\}$ using the k-th classification rule $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}_k^*)$. As before, let β_k be the number of individuals in $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ which are not drawn into $\underline{\mathbf{t}}_k^*$, α_k be the number of individuals in $\underline{\mathbf{t}}$ which are not drawn into $\underline{\mathbf{t}}_k^*$ and are misclassified by $W_k(\underline{\mathbf{x}}, \underline{\mathbf{t}}_k^*)$, and

$$\xi = \left(\sum_{k=1}^{b} \alpha_k \right) / \left(\sum_{k=1}^{b} \beta_k \right).$$

Finally, the balanced bootstrap estimator of overall actual error rate using mixture bootstrap samplings is given by

$$\hat{P}(MBB) = (1 - C_m)\hat{P}(R) + C_m \xi. \tag{4.1}$$

Here, $\hat{P}(R)$ is the *overall resubstitution error rate* (see Smith (1947)), and C_m represents the probability that the observation $\underline{\mathbf{x}}_{ij}$ in the original training sample $\underline{\mathbf{t}}$ is selected into the balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ obtained using mixture of bootstrapping (for $i = 1, 2$; $j = 1, 2, \ldots, n_i$; and $k = 1, 2, \ldots, b$).

The probability C_m for balanced bootstrap with mixture samplings, can be formulated as follows. Recall the resampling procedure in step (a) above. From this process, it is also obvious that the probability that an individual $\underline{\mathbf{x}}_{ij}$ in $\underline{\mathbf{t}}$ is included in the balanced bootstrap sample $\underline{\mathbf{t}}_k^*$ is the same as the probability that an element z in l is selected into l_k', z being any number from $\{1, \ldots, n\}$. For a particular value of z, L' will consist of b duplicates of z and $(bn - b)$ other numbers. Since L' is divided into b of l_k''s $(k = 1, 2, ..., b)$, each of size n, the probability that z is not selected into l', follows a hypergeometric model, and is given by

$$F(n, b) = \frac{\begin{pmatrix} b \\ 0 \end{pmatrix} \begin{pmatrix} nb - b \\ n \end{pmatrix}}{\begin{pmatrix} nb \\ n \end{pmatrix}} = \frac{(nb - b)!(nb - n)!}{(nb)!(nb - n - b)!}. \tag{4.2}$$

Hence, the probability that the individual $\underline{\mathbf{x}}_{ij}$ in $\underline{\mathbf{t}}$ is included in the balanced bootstrap sample $\underline{\mathbf{t}}_k^*$, which is the same as the probability of

an element z in l being selected into l', is given by

$$\Psi\left(n,b\right)=1-\frac{\left(\begin{array}{c}b\\0\end{array}\right)\left(\begin{array}{c}nb-b\\n\end{array}\right)}{\left(\begin{array}{c}nb\\n\end{array}\right)}=1-\frac{(nb-b)!(nb-n)!}{(nb)!(nb-n-b)!}. \quad (4.3)$$

So, the *finite* version of the balanced bootstrap estimator using mixture samplings can be defined (by replacing the coefficient C_m in equation (4.1) with $\Psi(n,b)$ in (4.3)) as

$$\hat{P}(FMB)=\left(1-\Psi(n,b)\right)\hat{P}(R)+\Psi(n,b)\xi. \quad (4.4)$$

To define the *infinite* version of this estimator, once again we have to find the convergence of the function $\Psi(n,b)$ in (4.3) for large values of n and b. Following the steps for obtaining the results in equation (3.10), we find that

$$\lim_{n\to\infty,b\to\infty}\Psi\left(n,b\right)=\lim_{n\to\infty,b\to\infty}\left(1-F\left(n,b\right)\right)=1-0.368=0.632. \quad (4.5)$$

Hence, the *infinite* version of the balanced bootstrap estimator using mixture samplings can be defined (by replacing the coefficient C_m in equation (4.1) with 0.632), as

$$\hat{P}(IMB)=0.368\hat{P}(R)+0.632\xi, \quad (4.6)$$

with IMB means *Infinite Mixture Balanced.*

5. Discussion

In this paper we have defined four versions of balanced bootstrap error rate estimators, namely *FSB* (Finite Separate Balanced), *ISB* (Infinite Separate Balanced), *FMB* (Finite Mixture Balanced), and *IMB* (Infinite Mixture Balanced). In practice, *ISB* and *IMB* seem easier to be computed since both these estimators have fixed coefficient 0.632.

However, either for small sample cases or for cases with small number of bootstrap samples, *FSB* or *FMB* seem to be more realistic estimators. To avoid the computations of the coefficients $\theta(n_1,n_2,b)$ in equation (3.8) and $\Psi(n,b)$ in equation (4.4), each time we use the estimators *FSB* and *FMB*, the values of function

$$F\left(n,b\right)=\frac{\left(\begin{array}{c}b\\0\end{array}\right)\left(\begin{array}{c}nb-b\\n\end{array}\right)}{\left(\begin{array}{c}nb\\n\end{array}\right)}$$

for various values of n and b, were first obtained (see Table 1).

Table 1. Values of function $F(n, b)$ given by equation (4.2) for various values of n and b.

Sample size (n)	Number of the bootstrap samples (b)							
	50	100	150	200	300	500	1000	$2000-\infty$
10	0.345	0.347	0.347	0.348	0.348	0.348	0.349	0.349
15	0.352	0.353	0.354	0.354	0.355	0.355	0.355	0.355
20	0.355	0.357	0.357	0.358	0.358	0.358	0.358	0.358
30	0.358	0.360	0.360	0.361	0.361	0.361	0.361	0.362
40	0.360	0.361	0.362	0.362	0.363	0.363	0.363	0.363
50	0.360	0.362	0.363	0.363	0.363	0.363	0.363	0.363
60	0.361	0.363	0.364	0.364	0.364	0.364	0.364	0.365
70	0.362	0.363	0.364	0.364	0.365	0.365	0.365	0.365
80	0.362	0.364	0.364	0.365	0.365	0.365	0.365	0.366
90	0.362	0.364	0.365	0.365	0.365	0.365	0.366	0.366
100	0.362	0.364	0.365	0.365	0.365	0.366	0.366	0.366
150	0.363	0.365	0.365	0.366	0.366	0.366	0.366	0.367
200	0.363	0.365	0.366	0.366	0.366	0.367	0.367	0.367
300	0.364	0.365	0.366	0.366	0.367	0.367	0.367	0.367
500	0.364	0.366	0.366	0.367	0.367	0.367	0.367	0.368
1000	0.364	0.366	0.366	0.367	0.367	0.367	0.368	0.368
$2000-\infty$	0.364	0.366	0.367	0.367	0.367	0.368	0.368	0.368

Thus, for given values of n_1, n_2, n, and b, the coefficient

$$\theta(n_1, n_2, b) = 1 - \left\{ \frac{n_1}{n_1 + n_2} F(n_1, b) + \frac{n_2}{n_1 + n_2} F(n_2, b) \right\}$$

and the coefficient $\Psi(n, b) = 1 - F(n, b)$, can be computed easily. For example, suppose that the first and the second training samples are of sizes $n = 20$ and $n = 30$, and the number of bootstrap samples is $b = 100$. For this case, the entire training sample is of size $n = n_1 + n_2 = 50$.

Then

$$\begin{aligned}\theta(20,30,100) &= 1-\left(\frac{20}{50}F(20,100)+\frac{30}{50}F(30,100)\right)\\ &= 1-((0.4)(0.357)+(0.6)(0.360))\\ &= 0.641,\end{aligned}$$

and $\Psi(50,100) = 1-F(50,100) = 1-0.362 = 0.638$. For this case,

$$\hat{P}(FSB) = 0.359\hat{P}(R)+0.641\xi$$

and

$$\hat{P}(FMB) = 0.362\hat{P}(R)+0.638\xi.$$

References

[1] Davison, A. C., Hinkley, D.V., and Schechtman, E. (1986). "Efficient Bootstrap Simulation," *Biometrika*, **73**, 555-566.

[2] Efron, B. (1983). "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, **78**, 316-331.

[3] Efron, B. (1990). "More Efficient Bootstrap Computations," *Journal of the American Statistical Association*, **85**, 79-89.

[4] Ganeshanandam, S., and Krzanowski, W.J. (1990). "Error-rate Estimation in Two-Group Discriminant Analysis Using The Linear Discriminant Function," *J. Statist. Comput. Simul.*, **36**, 157-175.

[5] Gleason, J. R. (1988). "Algorithm for Balanced Bootstrap Simulation," *The American Statistician*, **42**, 263-265.

[6] Graham, R. L. *et al*, (1990). "Balanced Design of Bootstrap Simulation," *Journal of the Royal Statistical Society, Ser. B*, **52**, 185-502.

[7] Hall, P., (1989a). "On Efficient Bootstrap Simulation," *Biometrika*, **76**, 613-617.

[8] Hall, P., (1989b). "Antithetic Resampling for the Bootstrap," *Biometrika*, **76**, 713-724.

[9] Hall, P., (1990). "Performance of Bootstrap Balanced Resampling in Distribution Function and Quantile Problems," *Probability Theory and Related Fields.*

[10] Hall, P. *et al*, (1989). "On Smoothing and the Bootstrap," *Annals of Statistics*, **17**, 692-704.

[11] Hinkley, D.V., and Shi, S. (1989). "Importance Sampling and the Nested Bootstrap," *Biometrika*, **76**, 435-446.

[12] Johns, M. V. (1988). "Importance Sampling for Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, **83**, 709-714.

[13] Mangku, I W. (1992). *Error Rate Estimation in Discriminan Analysis: Another Look at Bootstrap and Other Empirical Techniques.* Unpublished Master Thesis, Curtin University of Technology, Perth, Australia.

[14] Smith, C.A.B. (1947). "Some Examples of Discrimination," Annals of Eugenics, 13, 272-282.

[15] Snapinn, S.M., and Knoke, J.D. (1985). "An Evaluation of Smoothed Classification Error-Rate Estimators," *Technometrics*, **27**, 199-206.