A BOOTSTRAP INTERVAL ESTIMATOR FOR BAYES’ CLASSIFICATION ERROR

Chad M. Hawes and Carey E. Priebe

Johns Hopkins University
Department of Applied Mathematics and Statistics
Baltimore, MD 21218

ABSTRACT

Using a finite-length training set, we propose a new estimation approach suitable as an interval estimate of the Bayes-optimal classification error \( L^* \). We arrive at this estimate by constructing bootstrap training sets of varying size from the fixed, finite-length original training set. We assume a power-law decay curve for the unconditional error rate as a function of training sample size \( n \), and fit the bootstrap estimated unconditional error rate curve to this power-law form. Using a result from Devijver, we do this twice, once for the \( k \) nearest neighbor rule and again for Hellman’s \( (k, k') \) nearest neighbor rule with reject option, which gives a lower bound for \( L^* \). The result is an asymptotic interval estimate of \( L^* \) from a finite-length training sample. We apply our estimator to two classification examples, obtaining Bayes’ error estimates.

Index Terms— Error rate estimation, Bayes’ error, classification, bootstrap

1. INTRODUCTION

We propose a new error rate estimation approach suitable as an interval estimate of the Bayes-optimal probability of misclassification, from a finite sample. Given an observed, unlabeled random (feature) vector \( X \) to be classified, and independent, identically distributed training data \( \mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \) with \( (X_i, Y_i) : \Omega \to \mathbb{R}^d \times \{0, 1\} \) drawn from an unknown joint distribution \( F_{XY} \), the pattern recognition problem is to select a classification rule \( g : \mathbb{R}^d \to \{0, 1\} \) to predict the unknown class label \( Y \) with minimal misclassification error, where \( (X,Y) \) is distributed \( F_{XY} \) and independent of \( \mathcal{D}_n \). The performance of any classification rule cannot improve on the Bayes’ optimal error rate given by

\[
L^* \equiv \inf_{g: \mathbb{R}^d \to \{0,1\}} P\{g(X) \neq Y\}.
\]

Following the notational convention of [1], let us denote the finite sample conditional probability of error for the kNN-rule by \( L_n(k) = P\{g_{kNN}(X) \neq Y|\mathcal{D}_n\} \) and the unconditional error rate by \( \bar{L}_n(k) = E[L_n(k)] = P\{g_{kNN}(X) \neq Y\} \). We denote the asymptotic conditional and unconditional

\[
\hat{L}_\infty(k) = \lim_{n \to \infty} L_n(k) \quad \text{and} \quad \bar{L}_\infty(k) = \lim_{n \to \infty} \bar{L}_n(k),
\]

error rates by \( L_\infty(k) = \lim_{n \to \infty} L_n(k) \) and \( \bar{L}_\infty(k) = \lim_{n \to \infty} \bar{L}_n(k) \), respectively. Devijver [2] derives asymptotic upper and lower bounds on the Bayes-optimal classification error \( L^* \) using the k-nearest-neighbor (kNN) rule originally developed by Fix and Hodges [3] and the \( (k, k') \) NN-rule with reject option developed by Hellman [4]. Then in our notation, Devijver’s bounds are

\[
\bar{L}_\infty(k, k') \leq L^* \leq \tilde{L}_\infty(k),
\]

where \( \tilde{L}_\infty(k, k') \) is the asymptotic unconditional error rate of Hellman’s \( (k, k') \) NN-rule, where \( k' \) satisfies \( \frac{k + 1}{2} < k' \leq k \). All asymptotics here are for fixed \( k \) and \( k' \) as the training sample size \( n \to \infty \). This is great, we have tight bounds on the Bayes’ optimal error; unfortunately these are asymptotic bounds. In practice we never have at our disposal an infinite training sample, so that getting our hands on an estimate of \( \tilde{L}_\infty(k, k') \) and \( L_\infty(k) \) is non-trivial.

In practice, using Monte Carlo simulations, we often construct multiple simulated data sets for different sample sizes \( n \), evaluate an estimate of the statistical parameter of interest (e.g. detector false alarm rate in additive or multiplicative non-Gaussian noise), and construct a plot of the estimated statistical parameter as a function of the training sample size \( n \). The resulting curve can then be used to interpolate the expected system performance at desired sample sizes or to extrapolate the system performance if a parametric form for the estimated curve (e.g. linear, exponential) is recognized.

We extend Devijver’s large sample \( L^* \) kNN-rule bounding approach, using a finite sample, by this idea of estimating the error rate for various sample sizes. Our basic approach is to draw bootstrap training samples for different sample sizes \( n \) from the fixed, finite available training set \( \mathcal{D}_N \), construct an estimated error rate decay curve as a function of \( n \), and fit a parametric power-law form to the resulting curve. This process is performed for the standard kNN-rule to yield an estimate \( \tilde{L}_\infty(k) \) of \( \bar{L}_\infty(k) \) and then repeated using Hellman’s \( (k, k') \) NN-rule with reject option, giving an estimate \( \tilde{L}_\infty(k, k') \) for \( \bar{L}_\infty(k, k') \). The result is an asymptotic interval estimator \( \left(\tilde{L}_\infty(k, k'), \bar{L}_\infty(k)\right) \) for the Bayes-optimal classification error using a finite training sample size.
We caution that our approach will not work for all possible joint distributions $F_{XY}$, in keeping with the following statement by Devroye et al in [1]:

"However, it is impossible to estimate $L^*$ universally well: for any $n$, and any estimate of $L^*$ based upon the data sequence, there always exists a distribution of $(X, Y)$ for which the estimate is arbitrarily poor."

Thus there exists a restricted class of joint distributions $F_{XY}$ for which our $L^*$-estimation approach will work.

Our approach derives its ability to estimate the asymptotic unconditional error rate from a finite sample by the (strong) assumption of a three parameter power law form for the unconditional error rate decay curve:

$$L_n(k) = an^b + c.$$  \hspace{1cm} (2)

Implicitly this assumption places restrictions on the class of joint distributions $F_{XY}$ for which our technique will work. Given that our assumption holds, for $b<0$ as $n \to \infty$, we have $\hat{L}_n(k) \to c$, where $c = \hat{L}_\infty(k)$. Thus, fitting this power law model to the bootstrap sampling-constructed error rate decay curve yields an estimate for the kNN_rule’s asymptotic error rate $\hat{L}_\infty(k) = \hat{c}$.

There exists a rich literature on classification error rate estimation approaches, as the survey articles of Toussaint [5], Hand [6], and Schiavo and Hand [7] clearly show. In a similar vein, Fukunaga and Hummels [8] derive expressions for the bias between the finite sample error rate of the nearest neighbor rule and the Bayes’ optimal error. Integral to this approach are estimates of the finite sample error rate of nearest neighbor rules for different sample sizes to estimate this bias. For $d = 1$ and regularity conditions on the conditional densities, Cover [9] found that the nearest neighbor risk $L_n(1)$ converges to its asymptotic limit $L_\infty(1)$ according to $|L_n(1) - L_\infty(1)| \leq an^{-2}$, a power law of the form we advocate. Under suitable regularity conditions and for feature dimension $d$, Psaltis et al [10] found for the 1NN_rule and Snapp and Venkatesh [11] found for the kNN_rule that the finite sample unconditional error rate obeys the following asymptotic expansion:

$$\hat{L}_n(k) \sim \hat{L}_\infty(k) + \sum_{j=2}^{\infty} c_j n^{-j/d}.$$  \hspace{1.5cm}

For large $n$ the dominant term in the expansion is $cn^{-2/d}$, so that again a power-law form is found, dependent on the feature vector dimension $d$ and consistent with Cover’s INN result for $d = 1$. Thus there is a solid basis for the three parameter power-law form we espouse in (2). We have not found our bootstrap sampling, curve-fitting approach in the literature and believe this to be the first description of this particular error rate estimation approach.

The remainder of this paper is organized as follows. Section 2 provides a formal description of our finite sample interval estimator for the Bayes’ optimal error. Section 3 applies our estimation approach to two classification problems, and Section 4 summarizes our results and points to areas for future work.

2. ESTIMATION APPROACH

We assume the existence of independent, identically distributed (iid) training data $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1}^N$ and iid test data $\mathcal{M}_m = \{(X_i, Y_i)\}_{i=1}^M$, where we consider the binary classification problem with class label $Y$ taking values in $\{0, 1\}$ and feature vectors $X_i$ take values in $d$-dimensional Euclidean space. The samples $(X_i, Y_i)$ of $\mathcal{D}_N$ and $\mathcal{M}$ and $(X, Y)$ are drawn from an unknown joint distribution $F_{XY}$ and are mutually independent. We assume the number of available labeled training samples $N$ is fixed and finite.

For each $n_1 < n_2 < \cdots < n_m < N$, we construct bootstrap samples from the empirical joint distribution $\hat{F}_{XY}$, where $\hat{F}_{XY}$ places mass $1/N$ at each sample $(X_i, Y_i) \in \mathcal{D}_N$. For each $n_j$ we denote our bootstrap-constructed training set by $\mathcal{D}_{n_j} = \{(X_1, Y_1)^*, (X_2, Y_2)^*, \cdots, (X_n, Y_n)^*)\}$, where each element is drawn $(X_j, Y_j)^* \sim \hat{F}_{XY}$, with replacement. For this training set $\mathcal{D}_{n_j}$, we construct our kNN classifier $g_{n_j}(X; \mathcal{D}_{n_j})$, which we’ll denote by the simplified $g_{n_j}(X)$. Then we estimate its conditional error rate $\hat{L}_{n_j}(k)$ by the empirical errors evaluated on the test set $\mathcal{M}_m$:

$$\hat{L}_{n_j}(k) = \frac{1}{M} \sum_{i=1}^{M} I(g_{n_j}(X_i) \neq Y_i).$$  \hspace{1cm} (3)

Here $I(\cdot)$ is the indicator function and $(X_i, Y_i) \in \mathcal{M}_m$. At each desired sample size $n_j$, we construct $B$ bootstrap training sets of size $n_j$ and repeat this process, resulting in $B$ conditional error rate estimates $\{\hat{L}_{n_j}(k)\}_{j=1}^{B}$. From these $B$ conditional estimates, we construct an estimate of the unconditional error rate

$$\hat{L}_{n_j}(k) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{L}_{n_j}(k)\right)_b.$$  \hspace{1cm} (4)

This process is repeated for each $n_j$, resulting in estimated values of $\hat{L}_{n_j}(k)$ for the training sample sizes $n_1, n_2, \cdots, n_m$ and realizing an estimated error rate decay curve versus training sample size for the knn classifier rule. We now estimate the three parameters $a$, $b$, and $c$ in (2) by fitting the estimated $\{\hat{L}_{n_j}(k)\}_{j=1}^{m}$ values to this power-law decay, using a weighted nonlinear least squares approach ($\ell^2$ metric for the fit), with the weights given by the inverse of the estimated variance of the $B$ conditional error rate estimates at each $n_j$. This bootstrap variance weighting accounts for the inherent
Thus our final interval estimate for the Bayes’ optimal error is given by \( \left( \hat{L}_{\infty}(k), \hat{L}_{\infty}(k') \right) \).

Repeating this entire process using Hellman’s \((k, k')\) nearest neighbor rule with reject option and \( \left\lfloor \frac{k+1}{2} \right\rfloor < k' \leq k \) realizes the estimate \( \hat{L}_{\infty}(k, k') \). Finally, our interval estimate for the Bayes’ optimal error is given by \( \left( \hat{L}_{\infty}(k, k'), \hat{L}_{\infty}(k) \right) \).

3. EXAMPLES

We apply our estimation approach to two examples. The first is drawn from the joint distribution \( F_{XY}^{(PMH)} \) constructed in Section 3 of Priebe, Marchette, and Healy [12], where the feature vectors \( X \) take values in \( d = 6 \) dimensional Euclidean space and the known Bayes’ optimal classification error is \( L^* = 0.0653 \). We have available \( N = 2000 \) labeled samples from \( F_{XY}^{(PMH)} \) for our fixed training set \( \mathcal{D}_N \) and \( M = 2000 \) labeled samples in our test set \( \mathcal{F}_M \). We use a kNN classifier with \( k = 7 \) and the optimal nearest neighbor distance metric of Short and Fukunaga [13] adapted to the kNN rule. We adapt this optimal 1NN distance by computing the Euclidean distance from the current sample \( X \) (to be classified) to all training samples \( X_i \in \mathcal{D}_n^* \), select the \((2k+3)\) \( X \)-local training samples with the smallest Euclidean distance, and compute the optimal 1NN distance for these samples. Amongst these \((2k+3)\) samples, the closest \( k \) in the Short-Fukunaga 1NN distance metric are selected as the \( k \) nearest neighbors of \( X \). We use Hellman’s \((7,5)\) nearest neighbor rule with reject option.

We construct unconditional error rate estimates for bootstrap training set sample sizes from \( n_1 = 50 \) to \( n_20 = 1000 \) in steps of 50 with \( B = 200 \) bootstrap samples at each \( n_j \) and display the results in Figure 1(a). The symbols display the bootstrap-constructed error rate estimates for the kNN rule (circles) and the \((k, k')\) nearest neighbor rule with reject option (squares), while the solid lines show the power-law fits of (2). As explained above, we use weighted nonlinear least squares to fit the power-law curve to the bootstrap estimated error rates. For the Priebe-Marchette-Healy problem, our fitting procedure yielded the estimated power laws:

\[
\begin{align*}
\hat{L}_n(7) &= 7.84n^{-0.785} + 0.0678 \\
\hat{L}_n(7, 5) &= 22.42n^{-1.09} + 0.0565
\end{align*}
\]

Thus our final interval estimate for the Bayes’ optimal error is \( 0.0565 \leq L^* \leq 0.0678 \), which we see contains the true \( L^* = 0.0653 \).

The second example to which we apply our estimation approach is the Pima Indians diabetes data set from the University of California, Irvine machine learning repository [14], which consists of 768 training samples with the feature vectors drawn from \( d = 8 \) dimensional Euclidean space. We partitioned this data set once into \( N = 500 \) training samples in \( \mathcal{D}_N \) and \( M = 268 \) test samples in \( \mathcal{F}_M \). The Bayes’ optimal error for this dataset is unknown. A short literature search found the best reported classification error rate on the UCI Pima Indians diabetes data set to be 0.20 [15], meaning the Bayes’ optimal error will be less than this. We selected a kNN-rule with \( k = 11 \) and Hellman’s \((k, k')\) nearest neighbor rule with reject option using \( k = 11 \) and \( k' = 7 \). In both cases we use our adapted Short-Fukunaga distance measure. We constructed unconditional error rate estimates for bootstrap sample sizes from \( n_1 = 50 \) to \( n_{15} = 400 \) in steps of 25 with \( B = 200 \) bootstrap samples at each \( n_j \) and display the results in Figure 1(b).

We again used weighted nonlinear least squares to fit the
power-law curve to the bootstrap estimated error rates. For this Pima Indians diabetes problem, our fitting procedure yielded the estimated power laws:

\[
\hat{L}_n(11) = 0.31n^{-0.22} + 0.161 \\
\hat{L}_n(11, 7) = 0.59n^{-0.58} + 0.136
\]

Thus our final interval estimate for the Bayes’ optimal error is \(0.136 \leq L^* \leq 0.161\), which is below the best classification performance reported in the literature to date for this data set.

4. CONCLUSION

We have described a novel error rate estimation approach that enables us to form an asymptotic interval estimate for the Bayes’ optimal error using a final sample. We applied our bootstrap interval estimator to two classification problems available in the literature. For the Priebe-Marchette-Healy distribution [12], our interval estimate was shown to contain the true, known \(L^*\). On the UCI Pima Indians diabetes data set, our technique provided the first known interval estimate for the Bayes’ optimal error and we showed that this estimate is below the best reported classification performance to date on this data set, indicating that better classification rules may yet be found for this problem.

Our \(L^*-\)estimation technique will not work on all classification problems, but rather on some restricted set of distributions \(F_{XY}\). While we have explained and demonstrated our bootstrap interval estimation technique for the knn-rule, we know of no reason our technique would not work with other classification rules. As such, we believe our approach can be used to estimate the asymptotic error rate of a classification rule for some restricted subset of distributions. We do not yet know restrictions on the classes of distributions or types of classification rules for which our technique will work; these are areas for future work. Another area that warrants further research is how to select the parameters of Hellman’s \((k, k')\) nearest neighbor rule with reject option so as to form a reliable lower bound for \(L^*\).

5. REFERENCES


