Correspondence

Distribution-Free Performance Bounds with the Resubstitution Error Estimate

LUC P. DEVROYE AND TERRY J. WAGNER, MEMBER, IEEE

Abstract—**Probability inequalities are given for the deviation of the resubstitution error estimate from the unknown conditional probability of error. The inequalities are distribution free and can be applied to linear discrimination rules, to nearest neighbor rules with a reduced sample size, and to histogram rules.**

I. Introduction

In the nonparametric discrimination problem, one observes X, a random vector with values in $\mathbb{R}^d$, and wishes to estimate θ, a random variable known to take values in $\{1,\cdots,M\}$. All that is known about the distribution of (X,θ) is that which can be inferred from a sample $(X_1,\theta_1),\cdots,(X_n,\theta_n)$ drawn from the distribution of (X,θ). The sample, denoted by D_n, is assumed to be independent of (X,θ). If

$$\hat{\theta}=g(X,D_n)$$

denotes an estimate of θ from X and the sample, then

$$L_n=P\{\hat{\theta}\neq\theta|D_n\},$$

the probability of error given the sample, measures the performance of the estimate. Because the distribution of (X,θ) is unknown, there is no way of computing L_n from D_n. An immediate need then is to estimate it from the sample.

One of the oldest estimates of L_n is the resubstitution estimate (Toussaint [1])

$$\hat{L}_n=(1/n)\sum_1^n I_{\{\hat{\theta}_i\neq\theta_i\}} \tag{1}$$

where $\hat{\theta}_i=g(X_i,D_n)$, $1\leqslant i\leqslant n$, and $I_{\{\cdot\}}$ is the indicator function of the event $\{\cdot\}$. In this correspondence we obtain upper bounds for $P\{|\hat{L}_n-L_n|\geqslant\epsilon\}$ that do not depend on the distribution of (X,θ) and that apply to three types of discrimination rules or functions g:

1) linear discrimination rules;
2) condensed nearest neighbor rules;
3) histogram discrimination rules.

The existence of distribution-free bounds with the resubstitution estimate for linear discrimination rules was first noticed by Vapnik and Chervonenkis [2]. The bounds we present here for 1) improve the earlier ones given in Devroye and Wagner [3], while the ones we present for 2) and 3) are new.

II. Results

Let $\phi_1,\cdots,\phi_m$ be fixed functions from $\mathbb{R}^d$ to $\mathbb{R}$, and let

$$\omega_i=(\omega_{i0},\cdots,\omega_{im}),\qquad 1\leqslant i\leqslant M$$

Manuscript received June 28, 1977; revised July 3, 1978. This work was supported in part by the U.S. Air Force under Grant AFOSR 77-3385. This paper was presented at the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Troy, NY, June 6–8, 1977.

L. P. Devroye is with the School of Computer Science, McGill University, P.O. Box 6070 Station A, Montreal, PQ, Canada M3C 3G1.

T. J. Wagner was with the Department of Electrical Engineering, Rice University, Houston, TX. He is now with the Department of Electrical Engineering, University of Texas, Austin, TX 78712.

be any functions of D_n which take values in $\mathbb{R}^{m+1}$. Then the discrimination rule which takes $\hat{\theta}=i$, where i is the smallest integer for which

$$\sum_{j=1}^m \omega_{ij}(D_n)\phi_j(X)+\omega_{i0}(D_n) = \max_{1\leqslant k\leqslant M}\left\{\sum_{j=1}^m \omega_{kj}(D_n)\phi_j(X)+\omega_{k0}(D_n)\right\}, \tag{2}$$

is called a linear discrimination rule (see Duda and Hart [4]). For example, with $m=d$, $x=(x^1,\cdots,x^d)$, and $\phi_i(x)=x^i$, $1\leqslant i\leqslant d$, one obtains the ordinary linear discrimination rule. The following theorem is proved in Section IV.

Theorem 1: For every linear discrimination rule and $\epsilon>0$,

$$P\{|\hat{L}_n-L_n|\geqslant\epsilon\}\leqslant 4(4n/M)^{mM(M-1)}e^{-n\epsilon^2/8}. \tag{3}$$

Using the Borel–Cantelli lemma and Theorem 1, we see that for a given n and M, and uniformly over all linear discrimination rules, $|\hat{L}_n-L_n|\xrightarrow{n}0$ with probability one, a result of Glick [5]. In particular, picking $\omega_1,\cdots,\omega_M$ to minimize $\hat{L}_n$ is nearly equivalent for large n to minimizing L_n (Wagner [6]).

A consideration always present in nonparametric discrimination is how to implement rules derived from large amounts of data. Linear discrimination rules are of a fixed form and avoid this problem since the calculation of $\omega_1,\cdots,\omega_M$ need only be done once while the choice of $\phi_1,\cdots,\phi_m$ is usually dictated by computational simplicity. However, if one uses the nearest neighbor rule with D_n, a large n presents difficulties in that both storage requirements and computation times increase with n. In order to keep the implementation requirements within reason and still retain the intuitive appeal of the rule, various procedures for condensing or reducing the sample before the nearest neighbor rule is applied have been suggested beginning with Hart [7] and most recently by Ritter *et al.* [8] (see also Wilson [9], Tomek [10], Gates [11], and Wagner [12]). There seems to be ample evidence that a reduction, properly done, will improve the performance of the nearest neighbor rule over that obtained with the raw sample.

We assume below that the sequence which represents the condensed sample,

$$(Y_1,\xi_1),\cdots,(Y_K,\xi_K), \tag{4}$$

is obtained from D_n in any fashion where $K=K(D_n)$. The estimate $\hat{\theta}=\xi_j$ is then made whenever j is the smallest integer for which

$$\|X-Y_j\|=\min_{1\leqslant i\leqslant K}\|X-Y_i\|.$$

This, of course, is just the nearest neighbor rule used with (4). To use Theorem 2, it is assumed that $K(D_n)\leqslant k$, where k is known *a priori*. For example, one must continue to condense or reduce the sample until k points or less remain where k is chosen *a priori*.

Theorem 2: For any condensed sample with $K(D_n)\leqslant k$, the probability of error for the nearest neighbor rule with the condensed sample satisfies

$$P\{|\hat{L}_n-L_n|\geqslant\epsilon\}\leqslant 4(4n)^{d(k-1)k}e^{-n\epsilon^2/8}. \tag{5}$$

The condensed sample partitions $\mathbb{R}^d$ into K sets $A_1,\cdots,A_K$ associated with $Y_1,\cdots,Y_K$ (e.g., A_j is the set of points in $\mathbb{R}^d$ closer to Y_j than any other Y_i for $1 \leqslant i<j$ and as close to Y_j as any other Y_i for $j<i \leqslant n$). The partition here depends on D_n. If we fix the partition beforehand, we might expect to get even tighter bounds.

Let $A_1,\cdots,A_k$ be any *fixed* partition of $\mathbb{R}^d$, and let $\xi_1,\cdots,\xi_k$ be any $\{1,\cdots,M\}$-valued functions of D_n, where now $\hat{\theta}=\xi_j$ whenever $X \in A_j$. Such rules are here called histogram rules.

Theorem 3: For any $\epsilon>0$, for any fixed partition $A_1,\cdots,A_k$ of $\mathbb{R}^d$, and for *any* way of selecting $\xi_1,\cdots,\xi_k$ from D_n,

$$P\{|\hat{L}_n-L_n| \geqslant \epsilon\} \leqslant 2M^k e^{-2n\epsilon^2} \tag{6}$$

and

$$P\{|\hat{L}_n-L_n| \geqslant \epsilon\} \leqslant 4(1+(2n/k))^k e^{-n\epsilon^2/8}. \tag{7}$$

The condensed nearest neighbor and histogram discrimination rules are such that the estimate $\hat{\theta}$ takes at most k values and, as a consequence, the bounds (5) and (7) for $P\{|\hat{L}_n-L_n| \geqslant \epsilon\}$ do *not* depend upon M, which in fact may be infinite. Similar M-independent estimates exist for the linear discrimination rules if one decides to use at most k weight vectors ω_i instead of M.

III. Discussion

The bounds given in the three theorems, even for moderate n, can be useless for small M, d, m, and ϵ. The bound of Theorem 1 is, however, an asymptotic improvement over the one given in [3], while the bound of Theorem 2 and the second bound of Theorem 3 have the property that they do not depend on M. Also, unlike the bounds of Devroye and Wagner [3], all the ones presented here are of the form $\alpha_n \exp(-\beta n\epsilon^2)$, where $\beta>0$ is a constant and α_n is a function of n. One can conclude from this that

$$E\{(\hat{L}_n-L_n)^2\} \leqslant \frac{2}{\beta n}\log(\alpha_n+1),$$

that is, $E\{(\hat{L}_n-L_n)^2\}$ decreases as $1/n$ or $(\log n)/n$ for all the classes of rules considered in this paper. The proof is easy. Find ϵ_0 such that $\alpha_n \exp(-\beta n\epsilon_0^2)=\theta$, where $\theta>0$ is to be picked later. Then

$$\begin{aligned} E\{(\hat{L}_n-L_n)^2\} &= \int_0^\infty 2tP\{|\hat{L}_n-L_n|>t\}\,dt \\ &\leqslant \epsilon_0^2+2\int_{\epsilon_0}^\infty t\alpha_n e^{-\beta nt^2}\,dt \\ &= \frac{2}{\beta n}(\log(\alpha_n/\theta)+\theta) \end{aligned}$$

which is minimal for $\theta=1$.

In practice we often have to choose between several possible discrimination procedures. Past experience with similar data (medical, economic, administrative) can help in the selection, but there is no guarantee that *given the data* D_n, the selected discrimination method is best (has lowest probability of error L_n) among those under consideration. Assume now that for each procedure p in the collection $\mathcal{P}$, we compute an estimate $\hat{L}_n(p,D_n)$ of L_n and pick the one for which

$$\hat{L}_n(p^*,D_n)=\inf_{p\in\mathcal{P}} \hat{L}_n(p,D_n).$$

For the rules treated in this correspondence, the resubstitution estimate seems appropriate. However, while it is true that for all individual p, $P\{|\hat{L}_n-L_n| \geqslant \epsilon\} \leqslant \psi(p,n,\epsilon)$, the bound we have for p^* is

$$P\{|\hat{L}_n(p^*,D_n)-L_n(p^*,D_n)| \geqslant \epsilon\} \leqslant \sum_{p\in\mathcal{P}} \psi(p,n,\epsilon).$$

For example, if $\mathcal{P}=(p_1,\cdots,p_c)$ and p_i is a linear discrimination procedure with functions $\phi_1^i,\cdots,\phi_{m_i}^i$, then the selection method picks the best colllection of functions for the data D_n. The important point is that the inequality of Theorem 1 is not applicable to $\hat{L}_n(p^*,D_n)$ because the functions $\phi_1,\cdots,\phi_m$ depend on D_n. Fortunately, it is true that

$$P\{|\hat{L}_n(p^*,D_n)-L_n(p^*,D_n)| \geqslant \epsilon\} \leqslant \sum_{i=1}^{c} 4(4n/M)^{m_iM(M-1)}e^{-n\epsilon^2/8}.$$

It should be emphasized, however, that Theorem 1 is valid for *all* ways of picking the weight functions $\omega_1,\cdots,\omega_M$ from the data (e.g., to minimize $\hat{L}_n$) once $\phi_1,\cdots,\phi_m$ are fixed.

With the condensed nearest neighbor rule, we can compare l sequences $(Y_1,\xi_1),\cdots,(Y_k,\xi_k)$ of length k or less on the basis of $\hat{L}_n$. Regardless of how large l is, Theorem 2 applies to p^*, the rule with the seemingly best condensed sequence.

The inclusion of the nearest neighbor rule in $\mathcal{P}$ will force us to pick it for almost all D_n if our standard of comparison is the resubstitution estimate $\hat{L}_n$(i.e., $\hat{L}_n=0$, if X_1 has a density, independently of the value of L_n). This shows that for some rules other estimates of the probability of error must be used. In the case of the nearest neighbor rule, the deleted estimate (Cover [13], Rogers and Wagner [14]) seems to be the best candidate.

IV. Proofs

The key technique used in the proofs is due to Vapnik and Chervonenkis [15]. If $Y_1,Y_2,\cdots$ are independent random variables taking values in some abstract measure space $(\mathcal{Y},\mathcal{B})$ with $\nu(B)=P\{Y_i\in B\}$, for all $B\in\mathcal{B}$, $i=1,2,\cdots,$ then

$$P\Big\{\sup_{C\in\mathcal{C}} |\nu(C)-\nu_n(C)| \geqslant \epsilon\Big\} \leqslant 4s(\mathcal{C},2n)e^{-n\epsilon^2/8}$$

where

1) $\mathcal{C}$ is a subclass of $\mathcal{B}$,
2) $\nu_n(C)=(1/n)\Sigma_1^n I_{\{Y_i\in C\}}$,
3) $s(\mathcal{C},n)$ is the maximum over $y_1,\cdots,y_n$ in $\mathcal{Y}$ of the number of sets in $\{\{y_1,\cdots,y_n\}\cap C: C\in\mathcal{C}\}$.

The specific calculations for $s(\mathcal{C},n)$ that we shall need are the following. If $\mathcal{C}'$ represents the class obtained by intersecting ρ or less sets from $\mathcal{C}$ (or taking unions of ρ or less sets from $\mathcal{C}$), then

$$s(\mathcal{C}',n) \leqslant s(\mathcal{C},n)^\rho.$$

If $\mathcal{Y}$ is $\mathbb{R}^l$ and $\mathcal{C}$ is the class of linear half-spaces in $\mathbb{R}^l$, e.g., sets of the form

$$\Big\{x\in\mathbb{R}^l: \sum_1^l a_ix_i \leqslant a_0\Big\}$$

for some $a_0,\cdots,a_l$, then $s(\mathcal{C},n) \leqslant (2n)^l$. If the inequality used in the definition of $\mathcal{C}$ is made strict and/or reversed, the same bound can be used for $s(\mathcal{C},n)$. See Cover [16] for the details.

Proof of Theorem 1

Replacing $\mathcal{Y}$ with $R^m\times\{1,\cdots,M\}$, Y_i with (Φ_i,θ_i) (where $\Phi_i=(\phi_1(X_i),\cdots,\phi_m(X_i))$, $i=1,2,\cdots$), and $\mathcal{B}$ with the Borel subsets of $\mathbb{R}^m\times\{1,\cdots,M\}$, we see that

$$L_n = 1-\nu\left(\bigcup_1^M (A_i\times\{i\})\right)$$

$$\hat{L}_n = 1-\nu_n\left(\bigcup_1^M (A_i\times\{i\})\right)$$

where A_i is the set of all $y\in\mathbb{R}^m$ for which $\hat{\theta}=i$. Then

$$|\hat{L}_n-L_n| \leqslant \sup_{\mathcal{C}} |\nu_n(C)-\nu(C)|$$

where $\mathcal{C}$ is the class of sets of the form

$$\bigcup_1^M (A_i\times\{i\}).$$

Because A_i comes from an intersection of at most $M-1$ linear half-spaces, it is not difficult to see that

$$s(\mathcal{C},n) \leqslant \sup_{(n_1,\cdots,n_M):\ \Sigma n_j = n} \left(\prod_{j=1}^{M} (2n_j)^{m(M-1)} \right) \leqslant \left(\frac{2n}{M} \right)^{mM(M-1)},$$

and (3) follows.

Proof of Theorem 2

As in Theorem 1 we see that

$$L_n = 1 - \nu\left(\bigcup_{1}^{k} (B_i \times \{\xi_i\}) \right)$$

$$\hat{L}_n = 1 - \nu_n\left(\bigcup_{1}^{k} (B_i \times \{\xi_i\}) \right)$$

where $\xi_i \in \{1,\cdots,M\}$ and B_i is the set of $y \in \mathbb{R}^d$ closest to Y_i. Thus each B_i is the intersection of at most $k-1$ linear half-spaces in $\mathbb{R}^d$. If $\mathcal{C}'$ represents the class of sets from $\mathbb{R}^d \times \{1,\cdots,M\}$ of the form

$$\bigcup_{1}^{k} (B_i \times \{\xi_i\}),$$

then

$$s(\mathcal{C}',n) \leqslant s(\mathcal{C},n)^k$$

where $\mathcal{C}$ is the class of sets $B_1 \times \{\xi_1\}$. But

$$s(\mathcal{C},n) \leqslant \sup_{n_1,\cdots,n_M:\ \Sigma n_j = n} \left(\bigcup_{j=1}^{M} (2n_j)^{d(k-1)} \right) \leqslant (2n)^{d(k-1)}$$

where n_j is the number of points from $(y_1,\xi_1),\cdots,(y_n,\xi_n)$ which belong to $\mathbb{R}^d \times \{j\}$. Equation (5) now follows.

Proof of Theorem 3

For a fixed set C in $\mathbb{R}^d \times \{1,\cdots,M\}$, the inequality of Hoeffding [17] yields

$$P\{|\nu_n(C) - \nu(C)| \geqslant \epsilon\} \leqslant 2e^{-2n\epsilon^2}.$$

Because $A_1,\cdots,A_k$ are fixed in $\mathbb{R}^d$, the class $\mathcal{C}$ of sets we need to be concerned with are of the form

$$\bigcup_{1}^{M} (B_i \times \{i\})$$

where each B_j is a union of the A_i and the B_j partition $\mathbb{R}^d$. Because $\mathcal{C}$ has M^k members, we have (6) immediately from

$$P\{|\hat{L}_n - L_n| \geqslant \epsilon\} \leqslant P\left\{ \sup_{\mathcal{C}} |\nu_n(C) - \nu(C)| \geqslant \epsilon \right\} \leqslant 2M^k e^{-2n\epsilon^2}.$$

For the second part of the theorem, let $(x_1,y_1),\cdots,(x_n,y_n)$ be an arbitrary sequence, and let n_i be the number of j, $1 \leqslant j \leqslant M$, such that $A_i \times \{j\}$ contains at least one point from the sequence. Then the number of sets from

$$\{\{(x_1,y_1),\cdots,(x_n,y_n)\} \cap C:\quad C \in \mathcal{C}\}$$

is bounded by

$$\prod_{i=1}^{k} (1+n_i) \leqslant \left(\prod_{i=1}^{k} \frac{(1+n_i)}{k} \right)^k \leqslant \left(1 + \frac{n}{k}\right)^k.$$

Because the sequence was arbitrary, we conclude

$$s(\mathcal{C},n) \leqslant \left(1 + \frac{n}{k}\right)^k,$$

and (7) follows.

ACKNOWLEDGMENT

We thank Lois Feinholz for correcting a mistake in an earlier version of Theorem 3.

REFERENCES

[1] G. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 472–479, July 1974.
[2] V. N. Vapnik and A. Ya. Chervonenkis, "Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data," *Automat. Remote Contr.*, vol. 32, pp. 207–217, 1971.
[3] L. P. Devroye and T. J. Wagner, "A distribution-free performance bound in error estimation," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 586–587, Sept. 1976.
[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[5] N. Glick, "Sample-based classification procedures related to empiric distributions," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 454–461, July 1976.
[6] T. J. Wagner, "Another look at Φ-function pattern recognition," in *Proc. 2nd Asilomar Conf. Circuits and Systems*, 1968, pp. 442–443.
[7] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 515–516, May 1968.
[8] G. L. Ritter *et al.*, "An algorithm for a selective nearest neighbor decision rule," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 665–669, Nov. 1976.
[9] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 408–421, 1972.
[10] I. Tomek, "Two modifications of CNN," *IEEE Trans. Sys., Man, Cybern.*, vol. SMC-6, pp. 769–772, 1976.
[11] G. W. Gates, "The reduced nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 431–433, May 1972.
[12] T. J. Wagner, "Convergence of the edited nearest neighbor," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 696–697, Sept. 1973.
[13] T. M. Cover, "Learning in pattern recognition," in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic, 1968, pp. 111–132.
[14] W. H. Rogers and T. J. Wagner, "A finite sample distribution-free performance bound for local discrimination rules," *Ann. Stat.*, vol. 6, pp. 506–514, 1978.
[15] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of the relative frequencies of events to their probabilities," *Theory Prob. Appl.*, vol. 16, pp. 264–280, 1971.
[16] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-10, pp. 326–334, 1965.
[17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Ass.*, vol. 58, pp. 13–30, 1963.

Finite Memory Hypothesis Testing with Dependent Samples

JACK KOPLOWITZ

***Abstract*—Let $x_1, x_2, \cdots$ be a sequence of dependent random variables drawn from a probability measure P. Consider the hypothesis test H_0: $P = P_0$ versus H_1: $P = P_1$. It is shown that for a class of discrete valued processes, including Markov processes, the hypothesis test can be resolved with a three-state memory. The result is generalized to m-hypothesis tests which require $m+1$ states.**

I. INTRODUCTION

Let $x_1, x_2, \cdots$ be a sequence of random variables (rv) drawn from a probability measure P. We are interested in the hypothesis test H_0: $P = P_0$ versus H_1: $P = P_1$, under the constraint that the observations are summarized by a time-varying finite mem-

Manuscript received September 30, 1977; revised July 2, 1978. This work was supported by the National Science Foundation under Grants ENG 74-09817 and ENG 76-09374.

The author is with the Electrical and Computer Engineering Department, Clarkson College, Potsdam, NY 13676.