Proc. IFIP Congress 1974, Stockholm, Sweden, pp. 615–619, Aug. 5–10, 1974.
[8] C. E. Shannon, "Coding theorems for a discrete source with fidelity criterion," *IRE Nat. Convention Record*, Part 4, pp. 142–163, 1959.
[9] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.
[10] J. E. Savage, "The complexity of decoders," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 684–695, Nov. 1969.
[11] J. Pearl, "On the storage economy of inferential question-answering systems," *IEEE Trans. Syst., Man, and Cybern.*, vol. SMC-5, pp. 595–602, Nov. 1975.
[12] N. Pippenger, "Information theory and the complexity of switching networks," *Proc. 16th Annual Symp. on Foundations of Computer Science*, IEEE 75 CH 1003–34, Berkeley, CA, pp. 113–118, Oct. 13–15, 1975.
[13] J. Pearl, "On coding precedence relations with a pair-ordering fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 118–120, Jan. 1976.
[14] A. Chaitin, "A theory of program size formally identical to information theory," IBM, Yorktown Heights, NY, *REP. RC 4805*, Apr. 1974.

Correspondence

A Distribution-Free Performance Bound in Error Estimation

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

***Abstract*—It is shown that distribution-free confidence intervals can be placed about the resubstitution estimate of the probability of error of any linear discrimination procedure.**

I. Introduction

In the discrimination problem the statistician is given an *observation X*, a random vector taking values in $\boldsymbol{R}^d$, and wishes to estimate its *state* $\theta \in \{1,2\}$. The only knowledge that the statistician has of the distribution of X, given $\theta = i$, is that which can be inferred from a sample of size n_i drawn from F_i where

$$P[X \leq x \mid \theta = i] = F_i(x), \qquad i = 1,2. \tag{1}$$

The two samples, here called *data*, are denoted $X_1^1, \cdots, X_{n_1}^1$ and $X_1^2, \cdots, X_{n_2}^2$, respectively, and are assumed to be independent of X regardless of its state.

A discrimination procedure which has been frequently investigated in the past (see, for example, Duda and Hart [1, ch. 5]) is to estimate θ by $\hat{\theta}$ where

$$\hat{\theta} = \begin{cases} 1, & \text{if } w^tX \geq w_0 \\ 2, & \text{if } w^tX < w_0. \end{cases} \tag{2}$$

The vector $w^t = (w_1, \cdots, w_d)$ and the number w_0, called the weight vector and threshold weight, respectively, are chosen from the data. Regardless of what method is used to arrive at a weight vector and threshold weight, the statistician will always be interested in estimating

$$L_i = P[\hat{\theta} \neq i \mid X_1^1, \cdots, X^1{}_{n1}, X_1^2, \cdots, X_{n2}^2, \theta = i], \qquad i = 1,2,$$

a random variable whose value is just the frequency of errors when a large number of independent observations, all with state i, have their states estimated using (2).

The resubstitution estimates $\hat{L}_i$ of L_i are defined by

$$\hat{L}_2 = \frac{1}{n_2} \sum_1^{n_2} I_{[w^tX_j^2 \geq w_0]}$$

and

$$\hat{L}_1 = \frac{1}{n_1} \sum_1^{n_1} I_{[w^tX_j^1 < w_0]}.$$

These estimates have the appeal of being very simple to calculate once w and w_0 have been determined and, indeed, some procedures for finding w and w_0 involve the specific calculations above. For example, for a given $0 < \alpha < 1$, one may seek values w and w_0 such that, when $\hat{L}_1 \leq \alpha$, $\hat{L}_2$ is minimized.

The question that we address ourselves to here is: how much confidence can the statistician place in these estimates, that is, for a given $\epsilon > 0$, what is

$$P[|\hat{L}_i - L_i| < \epsilon]. \tag{3}$$

There is, of course, no way of calculating (3) since the distribution functions (1) are unknown. However, if μ_i denotes the measure on the Borel sets corresponding to F_i and $\hat{\mu}_i$ denotes the empirical measure on the Borel sets for $X_1^i, \cdots, X_{n_i}^i$ (e.g., $\hat{\mu}_i(A)$ is the proportion of the X with state i falling in the set A), then

$$P[|L_i - \hat{L}_i| \geq \epsilon] \leq P\left[\sup_{A \in \mathcal{C}_i} |\mu_i(A) - \hat{\mu}_i(A)| \geq \epsilon\right] \tag{4}$$

where $\mathcal{C}_i$ deontes the class of sets of the form $\{x: w^tx \geq w_0\}$, for $i = 2$, and $\{x: w^tx < w_0\}$, for $i = 1$. The random variable on the right in (4) is, in the one-dimensional case, essentially what is dealt with in the Glivenko–Cantelli theorem [2]. Indeed, for $d \geq 1$, Wolfowitz [2] showed that this random variable tends to zero with probability one as $n_i \to \infty$. While this gives the statistician some assurance that, for large n_i, his estimate of L_i will be close to the actual value uniformly in all procedures for determining w and w_0 (see Glick [3] for a thorough discussion of this point), he still falls short of getting a numerical grasp on (3).

Suppose now that $X_1, \cdots, X_n$ is a sample of size n drawn from the distribution function F. If μ denotes the measure corresponding to F and $\hat{\mu}$ denotes the empirical measure for $X_1, \cdots, X_n$, then Vapnik and Chervonenkis [4, theorem 2, p. 269] have shown that

$$P\left\{\sup_{A \in \mathcal{C}} |\mu(A) - \hat{\mu}(A)| \geq \epsilon\right\} \leq 4s(\mathcal{C}, 2n)e^{-n\epsilon^2/8}$$

where $\mathcal{C}$ is a class of Borel sets in $\boldsymbol{R}^d$ and $S(\mathcal{C}, n)$ is the maximum over $x_1, \cdots, x_n$ of the number of sets in $\{\{x_1, \cdots, x_n\} \cap A : A \in \mathcal{C}\}$. For the class of "half planes" that we are considering here (e.g., $\mathcal{C}_1$ or $\mathcal{C}_2$),

$$s(\mathcal{C}_1, n) = \sum_0^d \binom{n}{k} \leq n^d + 1, \qquad \text{if } n \geq d.$$

Manuscript received February 2, 1976. This work was supported in part by AFOSR Grant 72-2371.
The authors are with the Department of Electrical Engineering, University of Texas, Austin, TX 78712.

Applying these results to (4) yields

$$P[|\hat{L}_i - L_i| \geq \epsilon] \leq 4(1 + 2^d n_i^d)e^{-n_i\epsilon^2/8}, \qquad i = 1,2. \qquad (5)$$

The significance of (5) is that the statistician knows that

$$P[|\hat{L}_i - L_i| < \epsilon] \geq 1 - 4(1 + 2^d n_i^d)e^{-n_i\epsilon^2/8}, \qquad i = 1,2$$

regardless of F_1,F_2. By constraining his decision procedure to be linear, he can get a distribution-free performance bound with the resubstitution estimates $\hat{L}_i$ independently of the procedure used to find w and w_0. This generalizes the result stated in [5] for $d = 1$ and left there as an open question for $d > 1$.

II. Extensions

This result has easy extensions. Suppose the statistician decides to use a rule of the form:

$$\hat{\theta} = \begin{cases} 1, & \text{if } w^t\Phi(X) \geq w_0 \\ 2, & \text{if } w^t\Phi(X) < w_0 \end{cases}$$

where

$$\Phi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_m \end{pmatrix}$$

is a fixed vector of real-valued measurable functions defined on $\boldsymbol{R}^d$ with $w^t = (w_1, \cdots, w_m)$ and w_0 determined in some manner from the data. A distribution-free bound for

$$P[|\hat{L}_i - L_i| \geq \epsilon], \qquad i = 1,2,$$

can be obtained immediately by replacing X_j^i by $\Phi(X_j^i)$ so that m replaces d in (5). However, the Vapnik and Chervonenkis result allows a firmer bound if $s(\mathcal{C},n)$ can be computed, where $\mathcal{C}$ is the class of sets of the form $\{x \in \boldsymbol{R}^d: w^t\Phi(x) \geq w_0\}$ or $\{x \in \boldsymbol{R}^d: w^t\Phi(x) < w_0\}$. The early paper of Cover [6] contains some specific examples, including the important case where $w^t\Phi(x)$ is a polynomial of degree r in the components of x.

Suppose θ can now take values in $\{1, \cdots, M\}$ where

$$P[X \leq x/\theta = i] = F_i(x), \qquad 1 \leq i \leq M.$$

The data become the sequence

$$X_1^1, \cdots, X_{n_1}^1, \cdots, X_1^M, \cdots, X_{n_M}^M \qquad (6)$$

where $X_1^i, \cdots, X_{n_i}^i$ is a sample of size n_i drawn from F_i. The sequence (6) will be denoted simply by the vector D. The linear decision rule for M states is

$$\hat{\theta} = \text{smallest integer which achieves} \max_{1 \leq i \leq M} \{w_i^t X + w_{i0}\}, \quad (7)$$

where, as before, the weights and thresholds $w_1, w_{10}, \cdots, w_M, w_{M0}$ are determined in some manner from the data. If $L_i = P\{\hat{\theta} \neq i/D, \theta = i\}$, then its resubstitution estimate just counts the frequency of errors made by (7) on the sample $X_1^i, \cdots, X_{n_i}^i$. It is not very difficult to see that a distribution-free bound for this case is given by

$$P[|L_i - \hat{L}_i| \geq \epsilon] \leq 4(1 + 2^d n_i^d)^{M-1} e^{-n_i\epsilon^2/8}, \quad 1 \leq i \leq M. \quad (8)$$

Finally, we may assume, in some situations, that θ is a random variable taking values in $\{1, \cdots, M\}$ with an unknown distribution

$$P\{\theta = i\} = \pi_i, \qquad 1 \leq i \leq M. \qquad (9)$$

The data $(X_1,\theta_1), \cdots, (X_n,\theta_n)$ is now a sample of size n drawn from the distribution of (X,θ) which is determined from (1) and (9) while the random variable

$$L = P[\hat{\theta} \neq \theta \mid (X_1,\theta_1), \cdots, (X_n,\theta_n)] = \sum_1^M \pi_i L_i$$

is the probability of error for (7) with the statistician's data and his method of choosing the weights and thresholds. The resubstitution estimate of L becomes

$$\hat{L} = \frac{1}{n}\sum_1^n I_{[\hat{\theta}_i \neq \theta_i]} = \sum_1^M \frac{N_i}{n}\hat{L}_i = \sum_1^M \hat{\pi}_i \hat{L}_i$$

where N_i is the number of observations in the data with state i and $\hat{\pi}_i$ is the usual frequency estimate of π_i, $1 \leq i \leq M$. $\hat{L}$ is, of course, the frequency of errors made on the data with (7). For, $0 < \alpha < 1$,

$$\begin{aligned} &P[|\hat{L} - L| \geq \epsilon] \\ &\quad \leq P\left[\sup_i |\hat{\pi}_i - \pi_i| \geq \alpha\epsilon/M\right] \\ &\quad + P\left[\sup_i |\hat{\pi}_i - \pi_i| < \alpha\epsilon/M \text{ and } |\hat{L} - L| \geq \epsilon\right]. \end{aligned}$$

The second term above equals

$$\begin{aligned} &P\left[\sup_i |\hat{\pi}_i - \pi_i| < \alpha\epsilon/M \text{ and } \left|\sum_1^M \pi_i(\hat{L}_i - L_i)\right| \geq (1 - \alpha)\epsilon\right] \\ &\qquad \leq \sum_1^M P[|\hat{\pi}_i - \pi_i| \leq \alpha\epsilon/M \text{ and } |\hat{L}_i - L_i| \geq (1 - \alpha)\epsilon/M\pi_i]. \end{aligned}$$

Since $(1 - \alpha)\epsilon/M\pi_i \geq 1$ will yield a probability of zero in each term above, we consider only terms with

$$\pi_i \geq (1 - \alpha)\epsilon/M.$$

Then

$$[|\hat{\pi}_i - \pi_i| \leq \alpha\epsilon/M] \subseteq [N_i \geq n\epsilon(1 - 2\alpha)/M]$$

and, from (5) and assuming $1 - 2\alpha > 0$,

$$\begin{aligned} &P[|\hat{\pi}_i - \pi_i| \leq \alpha\epsilon/M \text{ and } |\hat{L}_i - L_i| \geq (1 - \alpha)\epsilon/M\pi_i] \\ &\qquad \leq 4(1 + 2^d(n\epsilon(1 - 2\alpha)/M)^d)^{M-1} e^{-n\epsilon^3(1-\alpha)^2(1-2\alpha)/8M^3} \end{aligned}$$

Using Hoeffding's inequality [7], we see that, for $0 < \alpha < \frac{1}{2}$,

$$\begin{aligned} &P[|L - \hat{L}| \geq \epsilon] \leq 2Me^{-2n\alpha^2\epsilon^2/M^2} \\ &\quad + 4M(1 + 2^d(n\epsilon(1 - 2\alpha)/M)^d)^{M-1} e^{-n\epsilon^3(1-\alpha)^2(1-2\alpha)/8M^3}. \quad (10) \end{aligned}$$

No attempt here has been made to find the tightest bound possible. The interest in (10), as stressed earlier, is that it works for *all* $\pi_1, \cdots, \pi_M, F_1, \cdots, F_M$ and *all* ways of choosing the weights and thresholds.

References

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[2] J. Wolfowitz, "Convergence of the empiric distribution function on half-spaces," in *Contributions to Probability and Statistics,* (Ed. by I. Olkin, *et al.*) Stanford, CA: Stanford Univ. Press, pp. 504–507, 1960.

[3] N. Glick, "Sample-based classification procedures related to empiric distribution functions," *IEEE Trans. Inform. Theory,* vol. IT-22, July 1976.

[4] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications XVI.* pp. 264–280, 1971.

[5] T. M. Cover and T. J. Wagner, "Topics in statistical pattern recognition," in *Communication and Cybernetics 10,* (Ed. by K. S. Fu) Berlin: Springer-Verlag, pp. 15–46, 1976.

[6] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electronic Computers,* vol. EC-14, pp. 326–334, 1965.

[7] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.,* vol. 58, pp. 13–30, 1963.