Distribution-Free Performance Bounds for Potential Function Rules

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

Abstract—In the discrimination problem the random variable θ, known to take values in $\{1,\cdots,M\}$, is estimated from the random vector X. All that is known about the joint distribution of (X,θ) is that which can be inferred from a sample $(X_1,\theta_1),\cdots,(X_n,\theta_n)$ of size n drawn from that distribution. A discrimination rule is any procedure which determines a decision $\hat{\theta}$ for θ from X and $(X_1,\theta_1),\cdots,(X_n,\theta_n)$. For rules which are determined by potential functions it is shown that the mean-square difference between the probability of error for the rule and its deleted estimate is bounded by $A/\sqrt{n}$ where A is an explicitly given constant depending only on M and the potential function. The $O(n^{-1/2})$ behavior is shown to be the best possible for one of the most commonly encountered rules of this type.

I. Introduction

Let $D_n=((X_1,\theta_1),\cdots,(X_n,\theta_n))$ be a sample of size n drawn from the distribution of (X,θ). If (X,θ) is independent of D_n then discrimination *rules* are ways of estimating the *state* θ from X and the sample, which assume only that X takes values in $\mathbb{R}^d$ and θ takes values in $\{1,\cdots,M\}$. Specifically, if $\hat{\theta}(n)=g_n(X,D_n)$ is the estimate of θ for the rule given by the function g_n: $\mathbb{R}^d\times(\mathbb{R}^d\times\{1,\cdots,M\})^n\to\{1,\cdots,M\}$, then

$$L_n=P\left[\hat{\theta}(n)\neq\theta\,|\,D_n\right]$$

is its probability of error for the given sample, and we are interested here in how one estimates L_n from D_n. (See Toussaint [1], Kanal [2], and Cover and Wagner [3] for surveys of the problem.)

If $\hat{L}_n$ is some estimate of L_n then one would like to know

$$\sup P\left[|\hat{L}_n-L_n|\geqslant\epsilon\right] \tag{1}$$

for $0<\epsilon<1$ where the supremum is taken over all distributions of (X,θ). As might be guessed, upper bounds to (1) seem to be the most for which one can hope. To be useful these bounds must go to zero with n, hopefully as fast as possible. For linear discrimination rules with the resubstitution error estimate, bounds to (1) have been found by Vapnik and Chervonenkis [4], Cover and Wagner [3], and Devroye and Wagner [5], [6]. For local rules (e.g., nearest neighbor rules) with the deleted error estimate, bounds to (1) have been found by Rogers and Wagner [7], Devroye and Wagner [8]. Bounds for (1) for other rules with the resubstitution error estimate may also be found in [8].

The class of rules which this paper considers may be described as follows. Let $K(x,y,\theta)$ be a nonnegative function defined on $\mathbb{R}^d\times\mathbb{R}^d\times\{1,\cdots,M\}$ and let

$$\sum_{i=1}^{n} K(X,X_i,\theta_i)I_{[\theta_i=j]}$$

be the vote for state j where $I_{[\cdot]}$ is the indicator function of the event $[\cdot]$. The estimate $\hat{\theta}(n)$ is taken to be the integer with the largest vote, or in the case of ties, the smallest integer from those tied. This class of rules is large enough to include the usual potential function methods where K is the potential function (Aizerman *et al.* [9], [10], Bashkirov *et al.* [11], [12]), histogram rules (Glick [13]), and two-step rules which use kernel density estimates with the same kernel widths [3]. Probably the simplest nontrivial rule from this class is obtained by putting

$$K(x,y,\theta)=I_{[\|x-y\|\leqslant r]}. \tag{2}$$

Then $\hat{\theta}(n)$ is just the integer with the highest frequency of occurrence from the integers θ_i with $\|X-X_i\|\leqslant r$, $1\leqslant i\leqslant n$. Mentioned first by Fix and Hodges [14], this rule is asymptotically optimal if r is allowed to vary with n. In particular, if L^* is the Bayes probability of error for estimating θ from X and if $r=r_n$ with

$$r_n\xrightarrow{n}0$$
$$nr_n^d\xrightarrow{n}\infty,$$

then $L_n\xrightarrow{n}L^*$ in probability regardless of the distribution of (X,θ) (Devroye and Wagner [15]).

One estimate of L_n, called the resubstitution estimate, is given by

$$L_n^R=\frac{1}{n}\sum_{1}^{n} I_{\left[\hat{\theta}_i\neq\theta_i\right]}$$

where $\hat{\theta}_i=g_n(X_i,D_n)$. Because (X_i,θ_i) is also in D_n, it is not surprising that L_n^R is frequently an optimistic estimate of L_n. For example, consider the simple rule with K given by (2). If r is less than $\|X_i-X_j\|$ for $1\leqslant i,j\leqslant n$, then L_n^R is always zero regardless of the value of L_n. From this it is not hard to see that (1), with this K and the resubstitution estimate, equals one for $0<\epsilon<1-1/M$. It appears then

Manuscript received October 25, 1978; revised February 6, 1979. T. Wagner was supported by the Air Force Office of Scientific Research under Grant 77-3385 and L. Devroye by the DoD Joint Services Electronics Program through the Air Force Office of Scientific Research under Contract F49620-77-C-0101.

L. P. Devroye is with the School of Computer Science, McGill University, 805 Sherbrooke St., West, Montreal, Canada, H3A ZK6.

T. J. Wagner is with the Department of Electrical Engineering, University of Texas at Austin, Austin, TX 78712.

0018-9448/79/0900-0601$00.75

that L_n^R is not a good estimate of L_n for the class of rules considered here.

One possible way to remove the optimistic tendency of L_n^R is to let $\hat{\theta}_i$ be the estimate from X_i and the sample with (X_i,θ_i) deleted, that is,

$$\hat{\theta}_i = g_{n-1}(X_i, D_{ni})$$

where

$$D_{ni} = ((X_1,\theta_1),\cdots,(X_{i-1},\theta_{i-1}),(X_{i+1},\theta_{i+1}),\cdots,(X_n,\theta_n)).$$

The resulting estimate is called the deleted estimate and is denoted L_n^D. To see how fast (1) might go to 0 with n for L_n^D, consider again the simple rule with K given by (2), let $M=2$, and let X and θ be independent with $P[\theta=1]=P[\theta=2]=\frac{1}{2}$. If r is bigger than the diameter of the support of X and n is even, $L_n^D=1$ whenever the number of $\theta_1,\cdots,\theta_n$ equal to one is $n/2$. Thus, for $0<\epsilon<\frac{1}{2}$,

$$P\left[|L_n^D - L_n| \geqslant \epsilon\right] \geqslant P\left[\sum_1^n I_{[\theta_i=1]} = n/2\right].$$

Using inequalities for factorials (Feller [16, p. 54]) we see that this last probability is greater than $1/\sqrt{2\pi n}$ so that (1) can go to zero no faster than $O(n^{-1/2})$ for the simple rule with K given by (2) and $0<\epsilon<\frac{1}{2}$. The main result of this paper is the following theorem.

Theorem: Let ρ^* be the smallest number $\rho \geqslant 1$ such that the range of K is contained in $\{0\}\cup[\alpha,\alpha\rho]$ for some $\alpha>0$. If no such ρ exists put $\rho^*=\infty$. Then (1) is bounded by

$$\sup E(L_n^D - L_n)^2/\epsilon^2$$

where

$$\sup E(L_n^D - L_n)^2 \leqslant \frac{1}{2n} + \frac{c\rho^*(M-1)}{\sqrt{n}}$$

and c is a constant independent of the underlying distribution and less than 24.0.

For the K of (2), $\rho^*=1$ so that (1) indeed goes to 0 as $O(n^{-1/2})$ for that simple rule. If K takes the values $0,1,2,\cdots,N$ then $\rho^*=N$, while if

$$K(x,y,\theta) = e^{-\|x-y\|^2/2\sigma^2}$$

or

$$K(x,y,\theta) = \begin{cases} T - \|x-y\|, & \|x-y\| \leqslant T \\ 0, & \text{elsewhere} \end{cases}$$

then $\rho^*=\infty$. We do not know if the above theorem can be extended to include these two interesting kernels.

II. Proofs

We begin by proving two lemmas. A rule is said to be *symmetric* if for each n the value of g_n does not depend on the order of the (X_i,θ_i) in D_n. In particular, the rules in the class defined above are all symmetric.

Lemma 1: For all symmetric rules

$$E(L_n^D - L_n)^2 \leqslant \frac{1}{2n} + 3E\left|I_{[g_n(X,D_n)\neq\theta]} - I_{[g_{n-1}(X,D_{n-1})\neq\theta]}\right| \leqslant \frac{1}{2n} + 3P\left[\hat{\theta}(n) \neq \hat{\theta}(n-1)\right].$$

Proof: Let

$$(X_t,\theta_t),(X_0,\theta_0),(X_1,\theta_1),\cdots,(X_n,\theta_n)$$

be independent identically distributed (i.i.d.) with the distribution of (X,θ), and for a,b,c contained in $\{t,0,1,2\}$ let

$$A_c^{a,b} = I_{[g_n(X_c,((X_a,\theta_a),(X_b,\theta_b),(X_3,\theta_3),\cdots,(X_n,\theta_n))\neq\theta_c]}$$

$$A_c^{a} = I_{[g_{n-1}(X_c,((X_a,\theta_a),(X_3,\theta_3),\cdots,(X_n,\theta_n))\neq\theta_c]}.$$

From Rogers and Wagner ([7, theorem 2.2]) we see that

$$E(L_n^D - L_n)^2 = \frac{1}{n}E(A_1^2(1-A_2^1)) + E\{A_1^{0t}A_2^{0t} - A_1^{02}A_2^0 + A_1^2A_2^1 - A_1^{t2}A_2^t\}. \quad (3)$$

Using Schwarz's inequality on the first term of (3), and noting that $(A_1^2)^2 = A_1^2$, $(1-A_2^1)^2 = 1 - A_2^1$ and $EA_2^1 = EA_2^1$, we see that this term is bounded by $1/2n$. For the second term of (3) we see from symmetry that it equals

$$\begin{aligned}
&E\{(A_1^{0t}-A_1^0)A_2^{0t} + (A_1^0 - A_1^{02})A_2^{0t} + (A_2^{0t}-A_2^0)A_1^{02} \\
&\quad + A_1^2(A_2^1 - A_2^{t1}) + (A_1^2 - A_1^{t2})A_2^{t1} + A_1^{t2}(A_2^{t1} - A_2^t)\} \\
&= E\{(A_1^{0t}-A_1^0)A_2^{0t} + (A_1^0 - A_1^{02})A_2^{0t} + (A_2^{0t}-A_2^0)A_1^{02} \\
&\quad + A_0^2(A_2^0 - A_2^{0t}) + (A_1^0 - A_1^{0t})A_0^{t1} + (A_1^{02} - A_1^0)A_2^{01}\} \\
&= E\{(A_1^{0t}-A_1^0)(A_2^{0t} - A_0^{t1}) + (A_1^0 - A_1^{02})(A_2^{0t} - A_2^{01}) \\
&\quad + (A_2^{0t} - A_2^0)(A_1^{02} - A_0^2)\} \\
&\leqslant 3E\{|A_0^{12} - A_0^2|\} \\
&= 3E\{|I_{[g_n(X,D_n)\neq\theta]} - I_{[g_{n-1}(X,D_{n-1})\neq\theta]}|\} \\
&\leqslant 3P[\hat{\theta}(n) \neq \hat{\theta}(n-1)]
\end{aligned}$$

and the lemma follows.

Lemma 2: Suppose $Y_1, Y_2,\cdots,$ are independent identically distributed with values in $[-1,-b]\cup\{0\}\cup[b,1]$ for some $0<b\leqslant 1$. Then

$$P(A) = P\left[\operatorname{sgn}\left(\sum_1^{n+1} Y_i\right) \neq \operatorname{sgn}\left(\sum_1^n Y_i\right)\right] \leqslant \frac{a}{b\sqrt{n+1}} \quad (4)$$

where $a<8.0$ and

$$\operatorname{sgn}(x) = \begin{cases} 1, & x>0, \\ 0, & x=0, \\ -1, & x<0. \end{cases}$$

Proof: If σ^2 denotes the variance of Y_1, then the Berry–Esseen inequality (Petrov [17, p. 111]) yields

$$\sup_x \left| P\left\{\sum_1^n (Y_i - EY_i) < \sigma\sqrt{n}\, x\right\} - \Phi(x)\right| \leqslant \frac{c_0}{\sigma^3}\frac{E|Y_1 - EY_1|^3}{\sqrt{n}} \leqslant \frac{2c_0}{\sigma\sqrt{n}} \quad (5)$$

where c_0 is a universal constant known to be less than 0.7975 (Van Beek [18]) and

$$\Phi(x) = \frac{1}{2\pi} = \int_{-\infty}^x e^{-t^2/2}\,dt.$$

From (5) we deduce that

$$P\left\{a' \leq \sum_1^n Y_i \leq b'\right\} \leq \frac{b'-a'}{\sqrt{2\pi}\,\sigma\sqrt{n}} + \frac{4c_0}{\sigma\sqrt{n}}. \qquad (6)$$

Additionally, $N=\sum_1^n I_{[Y_i \neq 0]}$, then (6) yields

$$P\left\{a' \leq \sum_1^n Y_i \leq b' | N\right\} \leq \frac{(b'-a')}{\sqrt{2\pi}\,\sigma_0\sqrt{N}} + \frac{4c_0}{\sigma_0\sqrt{N}} \qquad (7)$$

where $\sigma_0^2 = \mathrm{var}(Y_1 | Y_1 \neq 0)$. Letting $\lambda = EY_1$, $p=P\{Y_1 \neq 0\}$, $q=\lambda/p = E\{Y_1|Y_1\neq 0\}$ then two cases can occur.

1) If $\lambda^2 < p^2b^2/2$ and $Q=I_{[Y_{n+1}\neq 0]}$ then

$$\begin{aligned} &P\left\{\mathrm{sgn}\left(\sum_1^{n+1} Y_i\right) \neq \mathrm{sgn}\left(\sum_1^n Y_i\right) | N, Q\right\} \\ &\leq \frac{2Q}{\sigma_0\sqrt{2\pi}\,\sqrt{N}} + \frac{4c_0}{\sigma_0\sqrt{N}} \qquad (8) \\ &\leq \left(\frac{2Q}{\sqrt{\pi}} + 4c_0\sqrt{2}\right)/(b\sqrt{N}) \end{aligned}$$

since $\sigma_0^2 \geq b^2 - (\lambda/p)^2 > b^2/2$.

2) If $\lambda^2 \geq p^2b^2/2$ then the left side of (8) can be upper bounded by

$$\begin{aligned} &P\{S_n + Y_{n+1} - (N+Q)q \leq -(N+Q)q | N,Q\} \\ &\quad + P\{S_n - Nq \leq -Nq|N,Q\} \\ &\leq \frac{2\sigma_0^2}{Nq^2} \leq \frac{4}{Nb^2}, \end{aligned}$$

where $S_n = \sum_{i=1}^n Y_i$ for all n. Replacing c_0 with 0.7975, we see that

$$\frac{4}{Nb^2} < \frac{4c_0\sqrt{2}}{\sqrt{N}\,b}$$

whenever

$$\frac{4}{Nb^2} < 1$$

and, consequently,

$$P\{\mathrm{sgn}(S_{n+1}) \neq \mathrm{sgn}(S_n) | N,Q\} \leq \left(\frac{2Q}{\sqrt{\pi}} + 4c_0\sqrt{2}\right)/(b\sqrt{N}).$$

Now,

$$\begin{aligned} &P\{\mathrm{sgn}(S_{n+1}) \neq \mathrm{sgn}(S_n)\} \\ &= E\{P\{\mathrm{sgn}(S_{n+1}) \neq \mathrm{sgn}(S_n)|N,Q\} I_{[N\neq 0;\, Q\neq 0]}\} \\ &\quad + P\{N=0; Q\neq 0\} \\ &\leq E\left\{\left(\frac{2Q}{b\sqrt{\pi}\,\sqrt{N}} + \frac{4c_0\sqrt{2}}{b\sqrt{N}}\right) I_{[N\neq 0;\, Q\neq 0]}\right\} + (1-p)^n p \\ &= E\left\{\left(\frac{2p}{b\sqrt{\pi}\,\sqrt{N}} + \frac{4c_0\sqrt{2}\,p}{b\sqrt{N}}\right) I_{[N\neq 0]}\right\} + (1-p)^n p \\ &= \left(\frac{2p}{b\sqrt{\pi}} + \frac{4c_0\sqrt{2}\,p}{b}\right) \sum_{j=1}^n \left(\frac{j+1}{p(n+1)\sqrt{j}}\right)\binom{n+1}{j+1} \\ &\quad \cdot p^{j+1}(1-p)^{n-j} + (1-p)^n p \\ &\leq \left(\frac{2p}{b\sqrt{\pi}} + \frac{4c_0\sqrt{2}\,p}{b}\right) E\left\{\frac{\sqrt{2W}}{(n+1)p}\right\} + (1-p)^n p \end{aligned}$$

where W is an $(n+1,p)$ binomial random variable. Since $E\sqrt{W} \leq \sqrt{EW} = \sqrt{p(n+1)}$ and $(1-p)^n p \leq 0.5/(n+1)$, we obtain

$$\begin{aligned} &P\{\mathrm{sgn}(S_{n+1}) \neq \mathrm{sgn}(S_n)\} \\ &\leq \left(\frac{2}{\sqrt{\pi}} + 8c_0\right)\frac{1}{\sqrt{n+1}\,b} + \frac{0.5}{n+1} \\ &\leq \left(\frac{2}{\sqrt{\pi}} + 8c_0 + \frac{0.5}{\sqrt{n+1}}\right)\frac{1}{\sqrt{n+1}\,b} \end{aligned}$$

which proves the lemma.

Proof of Theorem: Let

$$Y_{ij} = K(x, X_i, 1) I_{[\theta_i = 1]} - K(x, X_i, j) I_{[\theta_i = j]}$$

for $1 \leq i \leq n$, $2 \leq j \leq M$. If

$$I_{[g_n(X,D_n)\neq 1]} \neq I_{[g_{n-1}(x,D_{n-1})\neq 1]}$$

then for some $2 \leq j \leq M$

$$\mathrm{sgn}\left(\sum_{i=1}^n Y_{ij}\right) \neq \mathrm{sgn}\left(\sum_{i=1}^{n-1} Y_{ij}\right).$$

But $Y_{ij},\cdots,Y_{nj}$ are i.i.d. with values which may be assumed to lie in $[-1,-1/\rho^*] \cup \{0\} \cup [1/\rho^*, 1]$. Thus

$$\begin{aligned} &E\left|I_{[\hat\theta(n)\neq\theta]} - I_{[\hat\theta(n-1)\neq\theta]}\right| \\ &\leq \operatorname*{ess\,sup}_{X,\theta} E\left\{\left|I_{[g_n(X,D_n)\neq\theta]} - I_{[g_{n-1}(X,D_{n-1})\neq\theta]}\right| | X,\theta\right\} \\ &\leq (M-1)\rho^* a/\sqrt{n} \end{aligned}$$

and the theorem follows from Lemma 1 and Chebychev's inequality.

III. Remarks

One would hope that the upper bound for sign changes in Lemma 2 could be improved by eliminating the dependence on b. It cannot. For example, if $X_1, X_2, \cdots$ are i.i.d. with

$$P[X_1 = -1] = 1/n,$$

and

$$P[X_1 = b] = 1 - (1/n),$$

where $b \in (1/n, 1/n - 1)$, then

$$\mathrm{sgn}\left(\sum_1^{n+1} X_i\right) \neq \mathrm{sgn}\left(\sum_1^n X_i\right)$$

if exactly one X_i equals one for $1 \leq i \leq n$. But the probability that exactly one X_i equals one is

$$\begin{aligned} &pn(1-p)^{n-1} \geq pn(1-p)^n \\ &\qquad \geq e^{-np/(1-p)} > 1/e^2, \qquad \text{for } n \geq 2. \end{aligned}$$

By extending Lemma 2 slightly, one can get a similar result to the above theorem for the holdout estimate. See Devroye and Wagner [19] for details.

Acknowledgment

We appreciate the careful attention of a reviewer who caught several mistakes in the original draft.

References

[1] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. 20, pp. 472–479, 1974.

[2] L. Kanal, "Patterns in pattern recognition: 1968–1974," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 697–722, 1974.

[3] T. M. Cover and T. J. Wagner, "Topics in statistical pattern recognition," *Commun. and Cybern.* vol. 10, pp. 15–46, 1975.

[4] 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.

[5] 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, 1976.

[6] ——, "Distribution-free performance bounds with the resubstitution error estimate," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 208–210, Mar. 1979.

[7] W. H. Rodgers and T. J. Wagner, "A finite-sample distribution-free performance bound for local discrimination rules," *Annals of Statistics*, vol. 6, pp. 506–514, 1978.

[8] L. P. Devroye and T. J. Wagner, "Distribution-free inequalities for the deleted and holdout error estimates," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 202–207, Mar. 1979.

[9] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automat. Remote Contr.*, vol. 25, pp. 917–936, 1964.

[10] ——, "The probability problem of pattern recognition learning and the method of potential functions," *Autom. Remote Contr.*, vol. 25, pp. 1307–1323, 1964.

[11] O. A. Bashkirov, E. M. Braverman and I. B. Muchnik, "Potential function algorithms for pattern recognition learning machines," *Automat. Remote Contr.*, vol. 25, pp. 692–695, 1964.

[12] E. M. Braverman and E. S. Pyatniskii, "Estimation of the rate of convergence of algorithms based on the potential functions method," *Automat. Remote Contr.*, vol. 27, pp. 80–100, 1966.

[13] N. Glick, "Sample-based multinomial classification," *Biometrics*, vol. 29, pp. 241–256, 1973.

[14] E. Fix and J. L. Hodges, "Discriminatory analysis: Nonparametric discrimination: Consistency properties," Rep. 4, Proj. no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[15] L. P. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," to appear in *Ann. Statist.*, May, 1980.

[16] W. Feller, *An Introduction to Probability Theory and its Applications*, New York: Wiley, vol. 1, 1968.

[17] V. V. Petrov, *Sums of Independent Random Variables*, New York: Springer–Verlag, 1975.

[18] P. Van Beek, "An application of Fourier methods to the problem of sharpening the Berry–Esseen inequality," *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, vol. 23, pp. 187–196, 1972.

[19] L. P. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Tech. Rep. no. 183, Electronics Research Center, University of Texas at Austin, 1976.