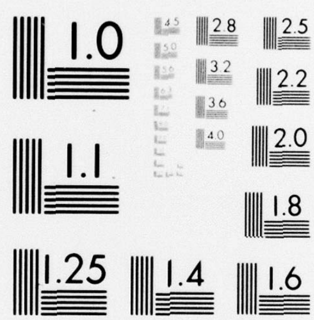
AD-A042 889 TEXAS UNIV AT AUSTIN DEPT OF ELECTRICAL ENGINEERING F/G 12/1
DISTRIBUTION-FREE PERFORMANCE BOUNDS WITH THE RESUBSTITUTION ER--ETC(U)
JUN 77 L P DEVROYE, T J WAGNER AF-AFOSR-2371-72
UNCLASSIFIED AFOSR-TR-77-0906 NL

1 OF 1
ADA042889

END
DATE
FILMED
9 -77
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR- 77-0906

Reprinted from the Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, Troy, NY, June 6-8, 1977.

AD A042889

DISTRIBUTION-FREE PERFORMANCE BOUNDS WITH THE RESUBSTITUTION ERROR ESTIMATE

Luc P. Devroye and T.J. Wagner*
Department of Electrical Engineering
The University of Texas
Austin, Texas 78712

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.

1. Introduction

The discrimination problem may be formulated as follows. The statistician collects data $(X_1,\theta_1),\ldots,(X_n,\theta_n)$, a sequence of independent identically distributed random vectors drawn from the distribution of (X,θ), a random vector independent of the data. For each $1 \le i \le n$, the observation X_i takes values in $\mathbb{R}^m$ and its state θ_i takes values in $\{1,\ldots,M\}$. The discrimination problem is that of estimating the state θ from the data and the observation X using procedures which do not require complete knowledge of the distribution of (X,θ). If $\hat\theta$ denotes the estimate, that is, $\hat\theta = g(X,V_n)$ where g is a Borel measurable $\{1,\ldots,M\}$-valued function of X and the data $V_n = (X_1,\theta_1,\ldots,X_n,\theta_n)$, then a measure of the performance of the procedure given the data is $L_n = P\{\hat\theta \ne \theta | V_n\}$, the conditional probability of error.

Since the distribution of (X,θ) is unknown, there is in general no way of computing L_n from the data. Using the data the statistician may try to estimate L_n by $\hat L_n$. A survey of estimation techniques can be found in Toussaint.[1] One of the oldest estimates is the resubstitution estimate

$$\hat L_n = n^{-1} \sum_{i=1}^{n} I_{\{\hat\theta_i \ne \theta_i\}}$$

where $\hat\theta_i = g(X_i,V_n)$, $1 \le i \le n$, are the estimates of the states of $X_1,\ldots,X_n$ with the given discrimination procedure, and where I is the indicator function.

In this paper we obtain upper-bounds for $P\{|L_n-\hat L_n| \ge \epsilon\}$ that do not depend upon the distribution of (X,θ), and that are applicable to three large classes of discrimination rules,

(i) the linear discrimination rules,
(ii) the nearest-neighbor rules with reduced sample size, and
(iii) the histogram decision rules.

The existence of distribution-free bounds with the resubstitution estimate for linear discrimination rules was first noticed by Vapnik and Chervonenkis.[10] The bounds for the class (i) improve the bounds given in Devroye and Wagner[2], while the results for the rules (ii) and (iii) are new. The possible existence of distribution-free bounds for (ii) was suggested to the authors by Dr. Penrod[16].

* This work was supported in part by the Air Force under Grant AFOSR 72-2371.

2. Main Results

Let $\phi_0, \phi_1,\ldots,\phi_{m'}$ be known measurable mappings from $\mathbb{R}^m$ to $\mathbb{R}$ where $m' \ge 1$ and $m \ge 1$, and $\phi_0 \equiv 1$. Let $w_0 = (w_{10},\ldots,w_{1m'}),\ldots, w_M = (w_{M0},\ldots,w_{Mm'})$ be Borel-measurable $\mathbb{R}^{m'+1}$-valued vector functions of the data V_n. Then, the rule which assigns the state $\hat\theta = j$ $(1 \le j \le M)$ to X whenever j is the first integer for which

$$\sum_{i=0}^{m'} w_{ji}(V_n)\, \phi_i(X) = \max_{1\le k\le M} \{\sum_{i=0}^{m'} w_{ki}(V_n)\, \phi_i(X)\}$$

is called a linear discrimination rule (see Duda and Hart[3] for a survey of the literature on linear discrimination). We emphasize that the $w_1,\ldots,w_M$ may be picked in an arbitrary fashion, using any method that can or cannot be found in the literature. The functions ϕ_i are picked in advance. The following bound is proved in the Appendix.

Theorem 1. For every $\epsilon > 0$ and for all linear discrimination rules with given $\phi_0,\phi_1,\ldots,\phi_{m'}$, the resubstitution estimate $\hat L_n$ satisfies

$$P\{|L_n-\hat L_n| \ge \epsilon\} \le 4M(1+(2n)^{m'})^{M-1} e^{-n\epsilon^2/8M^2} .$$

For the interesting case that M=2, we see that

$$P\{|L_n-\hat L_n| \ge \epsilon\} \le 8(1+(2n)^{m'})\, e^{-n\epsilon^2/32} .$$

Using the Borel-Cantelli lemma and Theorem 1, we see that for a given m' and M, and uniformly over all linear discrimination rules, $|L_n-\hat L_n| \overset{n}{\to} 0$ with probability one, a result due to Glick[11]. Thus, the statistician could pick the $w_1,\ldots,w_M$ that minimize the resubstitution estimate $\hat L_n$ because he knows from Theorem 1 that the corresponding probability of error L_n will be very close to $\hat L_n$, and that for large n, minimizing $\hat L_n$ is nearly equivalent to minimizing L_n (see Wagner[12]).

In the literature special attention has been given to the nearest-neighbor rule with a reduced number of observations where the reduction is a result of editing (see for instance Wilson[4]), condensing (Hart[5]) or any other operation (Tomek[6]). In general, we end up with $\mathbb{R}^m \times \{1,\ldots,M\}$-valued random vectors $(Y_1,\xi_1),\ldots,(Y_K,\xi_K)$ where K is an integer-valued random variable with $1 \le K \le k_n$. The (Y_i,ξ_i) and K may depend upon the data in an arbitrary fashion. A new observation X is assigned the state $\hat\theta = \xi_j$ whenever j is the smallest index for which

(See 1473)

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

Thus, $\hat{\theta}$ is the state of the nearest neighbor to X among $Y_1,\ldots,Y_K$. In the Appendix the following Theorem is proved.

Theorem 2. For every $\varepsilon > 0$, and for all the nearest-neighbor rules with reduced sample size, the resubstitution estimate $\hat{L}_n$ satisfies

$$P\{|L_n-\hat{L}_n| \ge \varepsilon\} \le 4M(1+(2n)^m)^{k_n-1} e^{-n\varepsilon^2/8k_n^2}$$

where k_n is an upper-bound on the reduced sample size.

We remark that this bound is independent of the distribution of (X,θ). The bound converges to 0 as n grows large provided that the sequence $k_1,k_2,\ldots$ is picked in such a way that $k_n^3 \log n/n \overset{n}{\to} 0$. It is clear that this bound is useless for the well-known nearest neighbor rule[7], that is, the rule with $K=k_n=n$ and $(Y_i,\xi_i) = (X_i,\theta_i)$, $1\le i\le n$. This was to be expected because the resubstitution estimate with the nearest-neighbor rule is overly optimistic. In fact, if the probability measure μ of X is absolutely continuous with respect to Lebesgue measure, then $\hat{L}_n = 0$ with probability one, no matter what value L_n takes.

Theorem 2 can be useful for reduced, selective, condensed or edited nearest-neighbor rules[4-6,13-14]. If k_n is a prespecified number of (Y_i,ξ_i)'s that are to be used in the new nearest-neighbor rule, then the statistician could compute $\hat{L}_n$ with some selected data $(X_{i_1},\theta_{i_1}),\ldots,(X_{i_{k_n}},\theta_{i_{k_n}})$ where $\{i_1,\ldots,i_{k_n}\}$ is a subset of $\{1,\ldots,n\}$, and decide to use that set of indices for which the resubstitution estimate is minimal. Using Theorem 2, we also know how much confidence we can put in our estimate $\hat{L}_n$ regardless of the selection procedure of the (Y_i,ξ_i) and without any knowledge of the distribution of (X,θ).

The (Y_i,ξ_i), $1\le i\le k_n$, partition $\mathbb{R}^m$ into k_n disjoint sets $A_1,\ldots,A_{k_n}$ where the state of X is estimated by $\hat{\theta} = \xi_j$ whenever X takes values in A_j (that is, X is closest to Y_j). The partition in this case depends on the data because the Y_i depend upon the data. For a given fixed partition of $\mathbb{R}^m$, we can expect to obtain tighter upper-bounds for $P\{|L_n-\hat{L}_n| \ge\varepsilon\}$ even if the partition is not generated by a reduced nearest-neighbor rule.

Let $A_1,\ldots,A_{k_n}$ be <u>any</u> fixed partition of $\mathbb{R}^m$ and let $\xi_1,\ldots,\xi_{k_n}$ be $\{1,\ldots,M\}$-valued random variables where, as before, ξ_j is the state assigned to X when-whenever X takes values in A_j. Such rules will be called *histogram decision rules*. We prove the following four distribution-free inequalities that are valid <u>no matter</u> how the ξ_i depend upon the data. The inequalities do not imply one another.

Theorem 3. For a given k_n-member partition of $\mathbb{R}^m$, for any way of specifying $\xi_1,\ldots,\xi_{k_n}$ from the data in a histogram decision rule, and for every $\varepsilon > 0$, the resubstitution estimate $\hat{L}_n$ satisfies

$$P\{|L_n-\hat{L}_n| \ge \varepsilon\} \le g_{ni} \;, \quad 1 \le i \le 4,$$

where

$$g_{n1} = 4\, k_n \operatorname{Min}(1+2n,M)\, e^{-n\varepsilon^2/8k_n^2}$$

$$g_{n2} = 2\, k_n\, M\, e^{-2n\varepsilon^2/M^2k_n^2}$$

$$g_{n3} = 4 \operatorname{Min}(M^{k_n}, 2^{2n}, (4n/k_n)^{k_n})\, e^{-n\varepsilon^2/8}$$

$$g_{n4} = 2\, M^{k_n}\, e^{-2n\varepsilon^2/M^2k_n} .$$

We note here that g_{n1} and g_{n3} are useful even if $M = \infty$ (i.e., the ξ_i and θ_i can take a countably infinite number of values). Clearly, all the g_{ni} are independent of the dimension m and the distribution of (X,θ), and $g_{n3} \overset{n}{\to} 0$ provided that $k_n/n \overset{n}{\to} 0$. If $k_n = \infty$, the bounds are not applicable, and, as we will see, the resubstitution estimate does not possess the distribution-free properties that it has with finite partitions of $\mathbb{R}^m$. *Assume that* $A_1,A_2,\ldots$ is a fixed countably infinite partition of $\mathbb{R}^m$. If the ξ_i are random variables that are independent of the data, then

$$P\{|L_n-\hat{L}_n| \ge \varepsilon\} \le 2e^{-2n\varepsilon^2} \tag{1}$$

for any $\varepsilon > 0$. However, such rules are impractical. The closest one can come to the Bayes rule with a fixed partition is to let $\xi_i = j$ if j is the smallest integer such that $N_{ij} = \max_{1<\ell<M} N_{i\ell}$ where N_{ij} is the number of (X_k,θ_k)'s with $X_k \in A_i$ and $\theta_k=j$. But even with *this obvious choice of* $\xi_1,\xi_2,\ldots$ *we see that for any m* and $M \ge 2$, there always exists a distribution of (X,θ) such that $|L_n-\hat{L}_n| \ge \frac{1}{2}$ with probability one. Indeed, assume that $M=2$, that $\theta=2$ with probability one, and that X takes values in each $A_1,\ldots,A_{2n}$ with equal probability $1/2n$. If the ξ_i are picked as described, then the resubstitution estimate $\hat{L}_n$ equals 0. Furthermore,

$$L_n = \sum_{i=1}^{2n} P\{X\in A_i\}\, I_{\{N_{i2}=0\}} \ge n/2n = 1/2.$$

This shows that even with the most obvious dependency of the ξ_i on the data, we will never be able to upper-bound $P\{|L_n-\hat{L}_n| \ge \varepsilon\}$ by an expression that decreases to 0 as n grows large, uniformly over all distributions of (X,θ).

3. Appendix

Proof of Theorem 1.

Let ν be the probability measure of (X,θ) where X takes values in $\mathbb{R}^{m'}$ and θ takes values in $\{1,\ldots,M\}$. It is clear that if ν_n is the empirical measure for $(X_1,\theta_1),\ldots,(X_n,\theta_n)$, and if $A_1,\ldots,A_M$ is the partition of $\mathbb{R}^{m'}$ that is generated by the linear discrimination rule (that is, A_i is the set on which we estimate the state of X by i), then

$$L_n = \sum_{i=1}^{M} \nu(A_i \times \{i\}^c)$$

and

$$\hat{L}_n = \sum_{i=1}^{M} \nu_n(A_i \times \{i\}^c) .$$

Thus,

$$|L_n - \hat{L}_n| = \Big|\sum_{i=1}^{M} (\nu(A_i\times\{i\}^c) - \nu_n(A_i\times\{i\}^c))\Big|$$

$$= \Big|\sum_{i=1}^{M} (\nu_n(A_i\times\{i\}) - \nu_n(A_i\times\{i\}))\Big|$$

$$\leq M \sup_{\substack{A\in\mathscr{A} \\ i\in\{1,\ldots,M\}}} |\nu_n(A\times\{i\}) - \nu(A\times\{i\})|$$

where $\mathscr{A}$ is the class of all sets that are intersections of (M-1) linear halfspaces of $\mathbb{R}^{m'}$. We recall that a linear halfspace of $\mathbb{R}^{m'}$ is a set of $x = (x^1,\ldots,x^{m'})$ for which $x^1a_1 + \ldots + x^{m'}a_{m'} \geq a_0$ or $x^1a_1 + \ldots + x^{m'}a_{m'} < a_0$ for some $(a_0,a_1,\ldots,a_{m'})\in\mathbb{R}^{m'+1}$. Thus, every $(a_0,a_1,\ldots,a_{m'})$ defines two linear halfspaces.

By an inequality of Vapnik and Chervonenkis[9],

$$P\{|L_n-\hat{L}_n| \geq \epsilon\}$$

$$\leq P\{\sup_{\substack{A\in\mathscr{A} \\ 1\leq i\leq M}} |\nu_n(A\times\{i\}) - \nu(A\times\{i\})| \geq \epsilon/M\}$$

$$\leq 4s(\mathscr{B},2n)e^{-n(\epsilon/M)^2/8}$$

where $\mathscr{B} = \mathscr{A}\times\{\{1\},\ldots,\{M\}\}$ and $s(\mathscr{B},n)$ is the maximum over all $(x_1,y_1),\ldots,(x_n,y_n)$ in $\mathbb{R}^{m'}\times\{1,\ldots,M\}$ of the number of different sets in $\{\{(x_1,y_1)\}\cup\ldots\cup\{(x_n,y_n)\} \cap B \,|\, B\in\mathscr{B}\}$. If $\mathscr{A}$ is the class of all linear halfspaces of $\mathbb{R}^{m'}$ and M=1, then $s(\mathscr{B},n) < 1+n^{m'}$ by a theorem of Cover[8] (see also Vapnik and Chervonenkis[9]). It is clear that if $\mathscr{A}$ is the class of all intersections of M^*-1 or less linear halfspaces and M=1, then $s(\mathscr{B},n) \leq (1+n^{m'})^{M^*-1}$. If M>1, then $s(\mathscr{B},n) \leq M(1+n^{m'})^{M^*-1}$. Indeed, if s_1 is the number of different sets in $\{\{x_1\}\cup\ldots\cup\{x_n\}\cap A \,|\, A\in\mathscr{A}\}$, then the number of different sets in $\{\{(x_1,y_1)\}\cup\ldots\cup\{(x_n,y_n)\}\cap B \,|\, B\in\mathscr{A}\times\{\{1\},\ldots,\{M\}\}\}$ is at most Ms_1. Thus we have shown that

$$P\{|L_n - \hat{L}_n| \geq \epsilon\} \leq 4M(1+(2n)^{m'})^{M-1} e^{-n\epsilon^2/8M^2} .$$

Q.E.D.

Proof of Theorem 2.

Let us use the notation of Theorem 1 where we let $A_1,\ldots,A_K$ be the partition of $\mathbb{R}^m$ that is generated by the nearest-neighbor rule with $(Y_1,\xi_1),\ldots,(Y_K,\xi_K)$ (i.e., A_i is the set on which we estimate the state of X by ξ_i and for which Y_i is the nearest neighbor to X among $Y_1,\ldots,Y_K$), then

$$L_n = \sum_{i=1}^{K} \nu(A_i\times\{\xi_i\}^c)$$

and

$$\hat{L}_n = \sum_{i=1}^{K} \nu_n(A_i\times\{\xi_i\}^c) .$$

Thus, arguing as in Theorem 1, we have

$$|L_n-\hat{L}_n| \leq K \sup_{\substack{A\in\mathscr{A} \\ 1\leq i\leq M}} |\nu_n(A\times\{i\}) - \nu(A\times\{i\})|$$

where $\mathscr{A}$ is the class of all sets that are intersections of (k_n-1) or less linear halfspaces of $\mathbb{R}^m$. Since $K \leq k_n$, we have by the argument of Theorem 1 that

$$P\{|L_n-\hat{L}_n|\geq\epsilon\} \leq 4M(1+(2n)^m)^{k_n-1} e^{-n\epsilon^2/8k_n^2} .$$

Q.E.D.

Proof of Theorem 3.

It is clear that

$$|L_n-\hat{L}_n| = \Big|\nu\Big(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\xi_\ell\}^c)\Big) - \nu_n\Big(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\xi_\ell\}^c)\Big)\Big|$$

$$= \Big|\nu\Big(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\xi_\ell\})\Big) - \nu_n\Big(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\xi_\ell\})\Big)\Big|$$

$$\leq \sum_{\ell=1}^{k_n} \sup_{1\leq i\leq M} |\nu(A_\ell\times\{i\}) - \nu_n(A_\ell\times\{i\})| .$$

Thus, if $\mathscr{C}_\ell$ is the class of sets of the form $A_\ell\times\{i\}$, $1\leq i\leq M$, then we know by an inequality of Vapnik and Chervonenkis[9] that

$$P\{|L_n-\hat{L}_n| \geq \epsilon\}$$

$$\leq k_n \sup_{1\leq\ell\leq k} P\{\sup_{C\in\mathscr{C}_\ell} |\nu(C)-\nu_n(C)| \geq \epsilon/k_n\}$$

$$\leq 4k_n\Big(\sup_{1\leq\ell\leq k_n} s(\mathscr{C}_\ell,2n)\Big)e^{-n(\epsilon/k_n)^2/8}$$

$$\leq 4k_n \operatorname{Min}(1+2n,M)\, e^{-n\epsilon^2/8k_n^2} .$$

Also,

$$P\{|L_n-\hat{L}_n| \ge \epsilon\}$$

$$\le P\left\{\sum_{\ell=1}^{k_n}\sum_{i=1}^{M} |\nu(A_\ell\times\{i\}) - \nu_n(A_\ell\times\{i\})| \ge \epsilon\right\}$$

$$\le k_n M \sup_{\substack{1\le i\le M\\ 1\le \ell\le k}} P\{|\nu(A_\ell\times\{i\}) - \nu_n(A_\ell\times\{i\})| \ge \epsilon/k_n M\}$$

$$\le 2k_n M e^{-2n\epsilon^2/M^2k_n^2}$$

by an inequality of Hoeffding[15]. Furthermore,

$$|L_n-\hat{L}_n|$$

$$\le \sup_{\substack{\text{all } (\lambda_1,\dots,\lambda_{k_n})\\ \text{from } \{1,\dots,M\}^{k_n}}} \left|\nu\left(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\lambda_\ell\})\right)-\nu_n\left(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\lambda_\ell\})\right)\right|$$

so that

$$P\{|L_n-\hat{L}_n|\ge\epsilon\} \le 4s(\mathscr{D}^*,2n)\, e^{-n\epsilon^2/8}$$

where $\mathscr{D}^*$ is the class of all sets of the form $\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\lambda_\ell\})$ where $(\lambda_1,\dots,\lambda_{k_n})\epsilon\mathscr{D} = \{1,\dots,M\}^{k_n}$. Clearly, $s(\mathscr{D}^*,2n) \le 2^{2n}$ for all k_n. However, if $k_n < 2n$, then $s(\mathscr{D}^*,2n) \le M^{k_n}$ and, in general, we must have that $s(\mathscr{D}^*,2n) \le 2^{k_n}(2n/k_n)^{k_n}$. This proves the inequality with g_{n3}.

Finally, notice that

$$P\{|L_n-\hat{L}_n| \ge \epsilon\}$$

$$\le \sum_{d\epsilon\mathscr{D}} P\left\{\left|\nu\left(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\lambda_\ell\})\right) - \nu_n\left(\bigcup_{\ell=1}^{k_n}(A_\ell\times\{\lambda_\ell\})\right)\right| \ge \epsilon/Mk_n\right\}$$

$$\le 2M^{k_n} e^{-2n\epsilon^2/M^{2k_n}} .$$

Q.E.D.

<u>Proof of (1).</u>

Inequality (1) is a corollary of Hoeffding's inequality[15] if we note that $|L_n-\hat{L}_n| = |\nu(C)-\nu_n(C)|$ where

$$C = \bigcup_{\ell=1}^{\infty} (A_\ell\times\{\xi_\ell\}^c) .$$

Q.E.D.

References

1. G. TOUSSAINT: "Bibliography on estimation of misclassification," <u>IEEE Transactions on Information Theory</u>, IT-20, pp. 472-479, 1974.
2. L.P. DEVROYE and T.J. WAGNER: "A distribution-free performance bound in error estimation," <u>IEEE Transactions on Information Theory</u>, IT-22, pp. 586-587, 1976.
3. R.O. DUDA and P.E. HART: <u>Pattern Classification and Scene Analysis</u>. New York: Wiley, 1973.
4. D.L. WILSON: "Asymptotic properties of nearest neighbor rules using edited data," <u>IEEE Transactions on Systems, Man and Cybernetics</u>, SMC-2, pp. 408-421, 1972.
5. P.E. HART: "The condensed nearest neighbor rule," <u>IEEE Transactions on Information Theory</u>, IT-14, pp. 515-516, 1968.
6. I. TOMEK: "Two modifications of CNN," <u>IEEE Transactions on Systems, Man and Cybernetics</u>, SMC-6, pp. 769-772, 1976.
7. T.M. COVER and P.E. HART: "Nearest neighbor pattern classification," <u>IEEE Transactions on Information Theory</u>, IT-13, pp. 21-27, 1967.
8. T.M. COVER: "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," <u>IEEE Transactions on Electronic Computers</u>, EC-10, pp. 326-334, 1965.
9. V.N. VAPNIK and A.Ya. CHERVONENKIS: "On the uniform convergence of the relative frequencies of events to their probabilities," <u>Theory of Probability and Its Applications</u>, vol. 16, pp. 264-280, 1971.
10. 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," <u>Automation and Remote Control</u>, vol. 32, pp. 207-217, 1971.
11. N. GLICK: "Sample-based classification procedures related to empiric distributions," <u>IEEE Transactions on Information Theory</u>, IT-22, pp. 454-461, 1976.
12. T.J. WAGNER: "Another look at Φ-function pattern recognition," in <u>Proceedings of the Second Asilomar Conference on Circuits and Systems</u>, pp. 442-443, 1968.
13. G.W. GATES: "The reduced nearest neighbor rule," <u>IEEE Transactions on Information Theory</u>, IT-18, pp. 431-433, 1972.
14. T.J. WAGNER: "Convergence of the edited nearest neighbor," <u>IEEE Transactions on Information Theory</u>, IT-19, pp. 696-697, 1973.
15. W. HOEFFDING: "Probability inequalities for sums of bounded random variables," <u>Journal of the American Statistical Association</u>, vol. 58, pp. 13-30, 1963.
16. C.S. PENROD: personal communication.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR-77-0906	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DISTRIBUTION-FREE PERFORMANCE BOUNDS WITH THE RESUBSTITUTION ERROR ESTIMATE		5. TYPE OF REPORT & PERIOD COVERED Interim
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L.P. Devroye ~~and~~ T.J. Wagner		8. CONTRACT OR GRANT NUMBER(s) AF-AFOSR-~~72~~-2371-72
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Texas Department of Electrical Engineering Austin, TX 78712		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5 A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332		12. REPORT DATE June 1977
		13. NUMBER OF PAGES 4 5p.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

PROCEEDINGS OF THE IEEE COMPUTER SOCIETY CONFERENCE ON PATTERN RECOGNITION AND IMAGE PROCESSING, Troy, NY, pp. 323-326, June 6-8, 1977.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Error Estimation; Discrimination

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

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.

401997

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED