On the Maximization of Divergence

DIMITRI KAZAKOS, MEMBER, IEEE

***Abstract*—An erroneous method for maximizing the projected divergence between two Gaussian multivariate hypotheses appeared in a recent paper. The correct solution is given.**

Introduction

Distance measures between statistical populations have found wide applicability in several diverse areas. Recent advances in source coding theory [1]–[3] and robust estimation [4] rely heavily on distance measures and their intrinsic properties. In addition, a large segment of the statistical pattern recognition literature deals with the problem of finding a linear transformation that maximizes some distance measure between classes [5]–[10].

In a recent paper [11], an erroneous method was used for finding the linear transformation that maximizes the divergence between two Gaussian populations. In the present correspondence we give the correct solution, which turns out to be a special case of the general result of [12].

Maximization of Divergence

Let $f_1(x)$, $f_2(x)$ be the probability density functions of the observation vector $x \in R^n$ under hypothesis H_1, H_2, respectively. The divergence J is defined as

$$J=\int_{R^n}[f_1(x)-f_2(x)]\log\left[f_1(x)f_2^{-1}(x)\right]dx. \tag{1}$$

If $f_i(x)$ are Gaussian with means M_i and covariance matrices Σ_i, $i=1, 2$, the divergence becomes

$$2J=(M_1-M_2)^t\left[\Sigma_1^{-1}+\Sigma_2^{-1}\right](M_1-M_2)+\operatorname{trace}\left[\Sigma_1^{-1}\Sigma_2+\Sigma_2^{-1}\Sigma_1-2I\right]. \tag{2}$$

Following the development of [11], let $\Sigma=\Sigma_1+\Sigma_2$, $R_i=P\Sigma_iP^t$, $i=1, 2$, and $M=P(M_1-M_2)$, where $P\Sigma P^t=I$. Let $Y=PX$ and $z=V^tY$, where V is an n-dimensional unit norm vector. Then the transformed divergence for the scalar random variable z and for the two hypotheses H_1, H_2 is

$$2J=\left[1+(V^tM)^2\right]\left[V^tR_1V-(V^tR_1V)^2\right]^{-1}-4 \tag{3}$$

under the condition

$$V^tV=1. \tag{4}$$

In order to find the value of V that maximizes J, we must seek the maximum of (3) under the constraint (4) because (3) is valid only for a unit norm V. In [11], the author ignores the unit norm constraint and proceeds to find the unconstrained extremal points of J by setting its gradient [11, (13)] equal to zero. As a result, his subsequent equations are in error. To correct this, let

$$L(V,\lambda)=J-2\lambda(V^tV-1). \tag{5}$$

Setting the gradient of L with respect to V equal to zero, we obtain

$$M(V^tM)V^tR_1V-R_1V\left(1+(V^tM)^2\right)=\lambda\left(V^tR_1V\right)^2\left(1-V^tR_1V\right)V. \tag{6}$$

Multiplying (6) by V^t we find

$$\lambda=-\left(1-V^tR_1V\right)^{-1}. \tag{7}$$

From (6) and (7)

$$M(V^tM)(V^tR_1V)=\left[\left(1+(V^tM)^2\right)R_1-(V^tR_1V)I\right]V. \tag{8}$$

The solution of (8) with respect to V provides the optimal V that maximizes J. The form of (8) is also derivable as a special case of Peterson and Mattson's theorem in [12].

Manuscript received November 29, 1976; revised November 22, 1977. This work was supported by the National Science Foundation under Grant ENG 76-20295.

The author is with the Department of Electrical Engineering, State University of New York at Buffalo, Amherst, NY 14260.

References

[1] R. M. Gray, D. L. Neyhoff, and D. S. Ornstein, "A generalization of Orstein's $\bar{d}$ distance with applications to information theory," *Ann. Prob.*, vol. 3, no. 3, pp. 478–491, Apr. 1975.

[2] R. M. Gray and L. D. Davisson, "Source coding without the ergodic assumption," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 502–516, July 1974.

[3] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.

[4] P. Papantoni-Kazakos, "Robustness in parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 2, pp. 223–230, Mar. 1977.

[5] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, pp. 52–60, Feb. 1967.

[6] T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 278–284, Apr. 1967.

[7] J. T. Tou and R. P. Heydorn, "Some approaches to optimum feature extraction," in *Computer and Information Sciences*, vol. 2, J. T. Tou, Ed. New York: Academic, 1967.

[8] T. L. Henderson and D. G. Lainiotis, "Comments on linear feature extraction," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 728–730, Nov. 1969.

[9] T. L. Henderson and D. G. Lainiotis, "Application of state variable techniques to optimal feature extraction—Multichannel analog data," *IEEE Trans. Inform. Theory*, vol. IT-16, no. 4, pp. 396–406, July 1970.

[10] P. Papantoni-Kazakos, "Some distance measures and their use in feature selection," Rice University Tech. Rep. #7611, Nov. 1976.

[11] C. B. Chittineni, "On the maximization of divergence in pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 590–592, Sept. 1976.

[12] D. W. Peterson and R. L. Mattson, "A method of finding linear discriminant functions for a class of performance criteria," *IEEE Trans Inform. Theory*, vol. IT-12, no. 3, pp. 380–387, July 1966.

On The Rate of Convergence of Nearest Neighbor Rules

LÁSZLÓ GYÖRFI

***Abstract*—The rate of convergence of the conditional error probabilities of the nearest neighbor rule and the *k*th nearest neighbor rule are investigated.**

Introduction

Let (X_0, θ_0), $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ be a sequence of independent identically distributed random vectors where X_j takes values in Euclidean d-space E^d and the possible values of its label θ_j are the integers $\{1, 2, \cdots, M\}$, $j=1, 2, \cdots, n$. The *a posteriori* probability functions will be denoted by

$$p_i(x)=P(\theta_j=i|X_j=x), \qquad i=1, 2, \cdots, M.$$

Given $Z^n=((X_1, \theta_1), \cdots, (X_n, \theta_n))$, we wish to estimate the label θ_0 of X_0. The nearest neighbor rule estimates θ_0 by the label of

Manuscript received March 15, 1976; revised October 28, 1977.

The author is with the Technical University of Budapest, 1111 Budapest, Stoczek u. 2. Hungary.

0018-9448/78/0700-0509$00.75 ©1978 IEEE

the NN of X_0, say $X'_{1,n}$, from the set $X_1,\cdots,X_n$ (see [1]) where the measure of closeness is defined by the Euclidean norm $|\cdot|$. If $\theta'_{1,n}$ denotes the label of $X'_{1,n}$, then the 1-NN decision is incorrect if $\theta_0 \neq \theta'_{1,n}$. For an integer $k<n$ the k-NN decision rule can be formulated as follows. Let $X'_{1,n}, X'_{2,n},\cdots,X'_{k,n}$ be the first, second,$\cdots$,kth NN of X_0 from the set $\{X_1,\cdots,X_n\}=X^n$. If for $i<j$ $|X_0-X_i|=|X_0-X_j|$ then X_i is closer to X_j by definition. Denote by $\theta'_{j,n}$ the label of $X'_{j,n}$. If L_i is the number of those labels that are equal to i, $1\leqslant i \leqslant M$, then the k-NN decision $\theta^*_{k,n}$ is equal to i if $L_i=\max\{L_1,L_2,\cdots,L_m\}$. If L_i and L_j are equal and maximal with $i<j$, then $\theta^*_{k,n}$ is taken equal to i. The k-NN decision makes an error if $\theta_0\neq\theta^*_{k,n}$.

We will make two assumptions.

A) The random variable X_0 has a continuous probability density f.

B) The *a posteriori* probabilities satisfy the weakened Lipschitz condition: there exists a Borel function $\boldsymbol{K}_i$ such that

$$|p_i(x)-p_i(y)|\leqslant \boldsymbol{K}_i(x)|x-y|, \qquad x,y\in E^d, i=1,2,\cdots,M.$$

T. Cover [2] has investigated the rate of convergence of $P(\theta_0\neq\theta'_{1,n})$ to R, the asymptotic error probability of the 1-NN rule. If $d=1$ and $M=2$, then he has proved that $|P(\theta_0\neq\theta'_{1,n})-R|=O(1/n^2)$ provided that the conditional densities of the random variable X_0 have third derivatives and these densities are bounded away from zero on their support sets. Under some mild conditions on the conditional densities, T. Wagner [3] and J. Fritz [4] have proved that $P(|P(\theta_0\neq\theta'_{1,n}|Z^n)-R|\geqslant E)$ converges exponentially fast.

Results

Using the notation of Cover and Hart [1], let

$$r(x)\triangleq 1-\sum_{i=1}^{M} p_i^2(x)$$

be the asymptotic conditional 1-NN error probability (or risk) for the point x, and let

$$r^*(x)\triangleq 1-\max_{1\leqslant i\leqslant M} p_i(x)$$

be the corresponding conditional Bayes risk. Let c_d be the Lebesgue measure of the unit sphere in E^d.

Theorem 1: With assumptions A, B we have for each $u>0$

$$\overline{\lim_{n\to\infty}}\, P(n^{1/d}|P(\theta_0\neq\theta'_{1,n}|X_0,X_1,\cdots,X_n)-r(X_0)|\geqslant u)\leqslant F(u), \tag{1}$$

where

$$F(u)=E\left[\exp\left\{-u^d\frac{c_d f(X_0)}{\left(\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)p_i(X_0)\right)^d}\right\}\right]. \tag{2}$$

The expression (2) has the disadvantage that it cannot be calculated because f, p_i, $\boldsymbol{K}_i$, $i=1,2,\cdots,M$ are unknown. However, (1) implies that

$$\lim_{n\to\infty} P(n^{\alpha}|P(\theta_0\neq\theta'_{1,n}|X_0,X^n)-r(X_0)|\geqslant u)=0$$

for $0<\alpha<1/d$ and $u>0$.

Theorem 2: With assumptions A, B we have for each $u>0$ and $k_n=[n^{2/(2+d)}]$

$$\overline{\lim_{n\to\infty}}\, P\left(\frac{n^{1/(2+d)}}{M+1}|P(\theta_0\neq\theta^*_{k_n,n}|X_0,Z^n)-r^*(X_0)|\geqslant u\right)\leqslant G(u), \tag{3}$$

where

$$G(u)=2Me^{-u^2}+P\left[\frac{1}{1+\frac{1}{d}}\cdot\frac{\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)}{(c_d f(X_0))^{1/d}}\geqslant u\right], \tag{4}$$

and $[x]$ denotes the integer part of x. If $k_n=[n^{\alpha}]$ where $0<\alpha<2/(2+d)$, then

$$\overline{\lim_{n\to\infty}}\, P\left(\frac{n^{\alpha/2}}{M+1}|P(\theta_0\neq\theta^*_{k_n,n}|X_0,Z^n)-r^*(X_0)|\geqslant u\right)\leqslant 2Me^{-u^2}. \tag{5}$$

Proofs

The proofs are mainly based on the limit distribution of $X'_{1,n}$ and on a weak law for $X'_{k_n,n}$.

Lemma 1: Under the condition A, for each $u>0$,

$$\lim_{n\to\infty} P(n^{(1/d)}|X_0-X'_{1,n}|\geqslant u|X_0)=\exp(-f(X_0)u^d c_d) \quad \text{a.s.} \tag{7}$$

Proof: By the definition of $X'_{1,n}$

$$P(n^{1/d}|X_0-X'_{1,n}|\geqslant\epsilon|X_0)=\left(1-Q\left[S\left(X_0,\frac{\epsilon}{n^{1/d}}\right)\right]\right)^n$$
$$=\exp\left(n\cdot\log\left(1-Q\left[S\left(X_0,\frac{\epsilon}{n^{1/d}}\right)\right]\right)\right), \tag{8}$$

where Q stands for the distribution of X_0 and $S(x,r)$ is the sphere of radius r centered at x. Because f is continuous,

$$Q\left(S\left(X_0,\frac{\epsilon}{n^{1/d}}\right)\right)=f(X_0)\frac{\epsilon^d c_d}{n}+\frac{\epsilon^d c_d}{n}o(1), \tag{9}$$

as $n\to\infty$. On the other hand, $Q(S(X_0,(\epsilon/n^{1/d})))>0$ a.s., and for each $0\leqslant z<1$

$$\log(1-z)=-z+O(z^2), \qquad z\to 0. \tag{10}$$

Therefore (8)–(10) imply (7).

Proof of Theorem 1: Cover and Hart [1] have shown that

$$P(\theta_0\neq\theta'_{1,n}|X_0,X^n)=1-\sum_{i=1}^{M}p_i(X_0)p_i(X'_{1,n}), \tag{11}$$

so that by condition B

$$|P(\theta_0\neq\theta'_{1,n}|X_0,X^n)-r(X_0)|\leqslant\sum_{i=1}^{M}p_i(X_0)|p_i(X_0)-p_i(X'_{1,n})|$$
$$\leqslant|X_0-X'_{1,n}|\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)p_i(X_0). \tag{12}$$

Applying the dominated convergence theorem, Theorem 1 follows from (12) and Lemma 1 since

$$\overline{\lim_{n\to\infty}}\, P(n^{1/d}|P(\theta_0\neq\theta'_{1,n}|X_0,X^n)-r(X_0)|\geqslant u)$$
$$\leqslant\overline{\lim_{n\to\infty}}\, E\left\{P\left(n^{1/d}|X'_{1,n}-X_0|\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)P_i(X_0)\geqslant u|X_0\right)\right\}$$
$$\leqslant E\left\{\lim_{n\to\infty}P\left(n^{1/d}|X'_{1,n}-X_0|\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)P_i(X_0)\geqslant u|X_0\right)\right\}$$
$$=E\left\{\exp\left[-\left\{\frac{u}{\sum_{i=1}^{M}\boldsymbol{K}_i(X_0)P_i(X_0)}\right\}^d c_d f(X_0)\right]\right\}. \tag{13}$$

Lemma 2: Let $\zeta_1, \zeta_2, \cdots, \zeta_n, \cdots$ be a sequence of independent identically distributed nonnegative random variables with a common continuous distribution function F. Denote by $\zeta^*_{1,n}, \cdots, \zeta^*_{n,n}$ the ordered sample of $\zeta_1, \cdots, \zeta_n$. Assume that for a real $r>0$, the limit

$$\lim_{z\to 0} \frac{F(z)}{z^r} \triangleq g_0 \tag{14}$$

exists and is positive. Then

$$\lim_{n\to\infty} \left(\frac{n}{k_n}\right)^{\frac{1}{r}} \frac{1}{k_n} \sum_{j=1}^{k_n} \zeta^*_{j,n} = \frac{1}{\left(1+\frac{1}{r}\right) g_0^{(1/r)}} \tag{15}$$

in probability.

Proof: Introduce the notation

$$F_r(z) \triangleq P(\zeta_1^r \leqslant z). \tag{17}$$

Then $-\log(1-F_r(\zeta_i^r))$, $i=1, 2, \cdots$, is a sequence of independent exponentially distributed random variables of parameter one. A. Renyi [9] has proved that

$$-\log\left(1-F_r(\zeta^{*r}_{j,n})\right) = \sum_{i=0}^{j-1} \frac{1}{n-i} \delta_{n,i} \tag{18}$$

where for fixed n $\{\delta_{n,0}, \delta_{n,1}, \cdots, \delta_{n,n}\}$ is a set of independent exponentially distributed random variables of parameter one. We show that

$$\left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \left(-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)\right)^{1/r} \to \frac{1}{1+\frac{1}{r}} \tag{19}$$

in mean square and therefore in probability. On the one hand (18) and $k_n/n \to 0$ imply that

$$\begin{aligned} &\lim_{n\to\infty} \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} E\left(-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)\right)^{1/r} \\ &= \lim_{n\to\infty} \frac{1}{k_n^{1+1/r}} \sum_{j=1}^{k_n} E\left(\sum_{i=0}^{j-1} \delta_{n,i}\right)^{1/r} \\ &= \lim_{n\to\infty} \frac{1}{k_n^{1+1/r}} \sum_{j=1}^{k_n} j^{\frac{1}{r}} E\left(\frac{1}{j}\sum_{i=0}^{j-1} \delta_{n,i}\right)^{1/r} = \frac{1}{1+\frac{1}{r}}, \end{aligned} \tag{20}$$

since $E[(1/j)\sum_{i=0}^{j-1}\delta_{n,i}]^{1/r}$ does not depend on n and tends to one if j tends to infinity. On the other hand by Jensen's inequality,

$$\begin{aligned} &\overline{\lim_{n\to\infty}}\, E\left[\left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \left(-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)\right)^{1/r}\right]^2 \\ &\leqslant \overline{\lim_{n\to\infty}}\, \frac{1}{k_n^{2+2/r}} \sum_{u=1}^{k_n} \sum_{v=1}^{k_n} E\left[\left(\sum_{i=0}^{u-1} \delta_{n,i}\right)^{1/r} \left(\sum_{l=0}^{v-1} \delta_{n,l}\right)^{1/r}\right] \\ &\leqslant \lim_{n\to\infty} \frac{1}{k_n^{2+2/r}} \sum_{u=1}^{k_n} \sum_{v=1}^{k_n} \left[\sum_{i=0}^{u-1} \sum_{l=0}^{v-1} E(\delta_{n,i}\delta_{n,l})\right]^{1/r} \\ &= \lim_{n\to\infty} \frac{1}{k_n^{2+2/r}} \sum_{u=1}^{k_n} \sum_{v=1}^{k_n} \left[\min(u,v) + u\cdot v\right]^{1/r} \\ &= \frac{1}{\left(1+\frac{1}{r}\right)^2}. \end{aligned} \tag{21}$$

Equations (20) and (21) imply (19). Let

$$h(x) \triangleq \sup_{0<y\leqslant x} \left| \frac{y}{\left(\frac{-\log(1-F_r(y^r))}{g_0}\right)^{1/r}} - 1 \right|.$$

Then by (14) and (17)

$$\lim_{x\to 0} h(x) = 0, \tag{22}$$

and

$$\begin{aligned} &\left| \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \zeta^*_{j,n} - \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \cdot \sum_{j=1}^{k_n} \left(\frac{-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)}{g_0}\right)^{1/r} \right| \\ &\leqslant \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \left(\frac{-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)}{g_0}\right)^{1/r} \cdot \left| \frac{\zeta^*_{j,n}}{\left(\frac{-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)}{g_0}\right)^{1/r}} - 1 \right| \\ &\leqslant \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \left(\frac{-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)}{g_0}\right)^{1/r} h(\zeta^*_{j,n}) \\ &\leqslant h(\zeta^*_{k_n,n}) \left(\frac{n}{k_n}\right)^{1/r} \frac{1}{k_n} \sum_{j=1}^{k_n} \left(\frac{-\log\left(1-F_r(\zeta^{*r}_{j,n})\right)}{g_0}\right)^{1/r} \to 0 \end{aligned}$$

in probability, since $\zeta^*_{k_n,n} \to 0$ a.s. (see [5]) and (19) and (22) are satisfied.

The proof of Theorem 2 needs a result of L. Györfi and Z. Györfi [5].

Lemma 3: For each $1/2>u>0$ and $1\leqslant i\leqslant M$

$$P\left(\left|\frac{L_i}{k_n} - \frac{1}{k_n}\sum_{j=1}^{k_n} P_i(X'_{j,n})\right| \geqslant u | X, X^n\right) \leqslant 2e^{-u^2 k_n}. \tag{26}$$

Lemma 4: Under conditions A and B, for each $u>0$, $k_n > [2u(M+1)]^2$,

$$\begin{aligned} &P\left(\frac{\sqrt{k_n}}{M+1} |P(\theta_0 \neq \theta^*_{k_n,n} | X_0, Z^n) - r^*(X_0)| \geqslant u | X_0, X^n\right) \\ &\leqslant 2Me^{-u^2} + \chi_{\{(\sum_{i=1}^M K_i(X_0)(1/\sqrt{k_n})\sum_{j=1}^{k_n}|X_0 - X'_{j,n}|) > u\}}. \end{aligned} \tag{27}$$

Proof: We show that

$$P(\theta_0 \neq \theta^*_{k,n} | X_0, Z^n) - r^*(X_0) \leqslant \sum_{i=1}^{M} \left|\frac{L_i}{k_n} - P_i(X_0)\right|.$$

Let us denote by $\tilde{A}_1, \cdots, \tilde{A}_M$ the partition of E^d for the decision $\theta^*_{k,n}$ so that $\tilde{A}_i = \{\theta^*_{k,n} = i\}$, $i=1, \cdots, M$, and by $A_1, \cdots, A_M$ the partition of the Bayesian decision. By the definition of k_n-NN decision, $(L_i/k_n) - (L_j/k_n) \geqslant 0$ on $\tilde{A}_i$, for each $i, j = 1, \cdots, M$, and

$$\begin{aligned} P(\theta_0 \neq \theta^*_{k,n} | X_0, Z^n) &= 1 - \sum_{i=1}^{M} P(\theta_0 = i, \theta^*_{k,n} = i | X_0, Z^n) \\ &= 1 - \sum_{i=1}^{M} \chi_{\tilde{A}_i}(X_0) P_i(X_0). \end{aligned}$$

Therefore

$$\begin{aligned} &P(\theta_0 \neq \theta^*_{k,n} | X_0, Z^n) - r^*(X_0) \\ &= \sum_{i=1}^{M} \chi_{A_i}(X_0) P_i(X_0) - \sum_{i=1}^{M} \chi_{\tilde{A}_i}(X_0) P_i(X_0) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{M} \chi_{A_i \cap \tilde{A}_j}(X_0) (P_i(X_0) - P_j(X_0)). \end{aligned} \tag{28}$$

Since $P_i(X_0) \geq P_j(X_0)$ on A_i and $L_i \geq L_j$ on $\tilde{A}_i$,

$$\begin{aligned} &P(\theta_0 \neq \theta^*_{k,n}|X_0, Z^n) - r^*(X_0) \\ &= \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \chi_{(A_i \cap \tilde{A}_j) \cup (A_j \cap \tilde{A}_i)}(X_0)|P_i(X_0) - P_j(X_0)| \\ &\leq \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \chi_{(A_i \cap \tilde{A}_j) \cup (A_j \cap \tilde{A}_i)}(X_0) \\ &\quad \cdot \left(\left| P_i(X_0) - \frac{L_i}{k_n} \right| + \left| P_j(X_0) - \frac{L_j}{k_n} \right| \right) \\ &\leq \sum_{i=1}^{M} \left| P_i(X_0) - \frac{L_i}{k_n} \right|. \end{aligned} \tag{29}$$

Applying (29) and condition B

$$\begin{aligned} &P(\theta_0 \neq \theta^*_{k,n}|X_0, Z^n) - r^*(X_0) \\ &\leq \sum_{i=1}^{M} \left| \frac{L_i}{k_n} - \frac{1}{k_n} \sum_{j=1}^{k_n} P_i(X'_{j,n}) \right| \\ &\quad + \left(\sum_{i=1}^{M} K_i(X_0) \right) \left(\frac{1}{k_n} \sum_{j=1}^{k_n} |X_0 - X'_{j,n}| \right). \end{aligned} \tag{30}$$

If $0 < u < 1/2$, then by (30) and Lemma 3, we get

$$\begin{aligned} &P(|P(\theta_0 \neq \theta^*_{k_n,n}|X_0, Z^n) - r^*(X_0)| \geq u | X_0, X^n) \\ &\leq P\Bigg(\bigcup_{i=1}^{M} \left\{ \left| \frac{l_i}{k_n} - \frac{1}{k_n} \sum_{i=1}^{k_n} P_i(X'_{j,n}) \right| \geq \frac{u}{M+1} \right\} \\ &\quad \cup \left\{ \left(\sum_{i=1}^{M} K_i(X_0) \right) \left(\frac{1}{k_n} \sum_{j=1}^{k_n} |X_0 - X'_{j,n}| \right) \geq \frac{u}{M+1} \right\} \Bigg| X_0, X^n \Bigg) \\ &\leq 2 \cdot M e^{-(u/(M+1))^2 k_n} \\ &\quad + \chi_{\{(\sum_{i=1}^{M} K_i(X_0))((1/k_n)\sum_{j=1}^{k_n}|X_0 - X'_{j,n}|) \geq u/(M+1)\}}. \end{aligned} \tag{31}$$

Under the condition of Lemma 4, u can be replaced by $u(M+1)/\sqrt{k_n}$, and then (31) implies (27).

Proof of Theorem 2: In case $\zeta_i = |X_0 - X_i|$, $r \triangleq d$ Lemma 2 implies that for $k_n \to \infty$, $k_n/n \to 0$ and $\epsilon > 0$

$$\lim_{n \to \infty} P \left\{ \left| \left(\frac{n}{k_n} \right)^{1/d} \frac{1}{k_n} \sum_{j=1}^{k_n} |X_0 - X'_{j,n}| - \frac{1}{\left(1 + \frac{1}{d}\right)(c_d f(X_0))^{1/d}} \right| > \epsilon \Bigg| X_0 \right\} = 0.$$

Consequently for each z

$$\overline{\lim_{n \to \infty}} P\left(\left(\frac{n}{k_n} \right)^{1/d} \frac{1}{k_n} \sum_{j=1}^{k_n} |X_0 - X'_{j,n}| \geq z \Big| X_0 \right) \leq \chi_{\{1/(1+(1/d))(c_d f(X_0))^{1/d} \geq z\}}. \tag{32}$$

Therefore (32), Lemma 4, and $k_n = [n^\alpha]$ imply Theorem 2.

Acknowledgment

The author is grateful to Professor T. J. Wagner, who detected a basic error in the original draft during the revision.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 21–27, Jan. 1967.

[2] T. M. Cover, "Rates of convergence for nearest neighbor procedures," *Proc. Hawaii Int. Conf. on System Sciences*, Honolulu, Hawaii, 1968.

[3] T. J. Wagner, "Convergence of the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566–571, Sept. 1971.

[4] J. Fritz, "Distribution-free exponential error bound for nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 552–557, Sept. 1975.

[5] L. Györfi and Z. Györfi, "On the nonparametric estimate of *a posteriori* probabilities of simple statistical hypotheses," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Amsterdam: North-Holland, 1977, pp. 298–308.

[6] A. Rényi, "Wahrscheinlichkeitsrechnung," (VIII. §9.) *VEB Deutscher Verlag der Wissenschaften*, Berlin, 1962.

An Upper Bound on the Asymptotic Error Probability of the k-Nearest Neighbor Rule for Multiple Classes

LÁSZLÓ GYÖRFI AND ZOLTÁN GYÖRFI

***Abstract*—If R_k denotes the asymptotic error probability of the k-nearest neighbor rule for M classes and R^* denotes the Bayes probability of error, then conditions are given that yield $R_k - R^* \leq \sqrt{MR_1/k}$.**

Introduction

If k is a fixed odd number, then Cover and Hart [1] have calculated the asymptotic error probability of the k-NN rule for the two-class pattern classification problem. Our goal is to give an upper bound for the error probability of the k-NN for arbitrary fixed k and an arbitrary number of classes. We investigate the conditional error probability $L_{k,n}$ and the asymptotic error probability R_k of the k-NN rule. Wagner [5] and Fritz [6] dealt with the almost sure convergence of $L_{1,n}$ in the case when the sample space is Euclidean space and the observation is nonatomic. If the *a posteriori* probability functions satisfy the Cover–Hart condition [1] (see the theorem), if $k/n \xrightarrow{n} 0$, and if, for each $\epsilon > 0$, $\sum_{n=1}^{\infty} e^{-\epsilon k_n} < +\infty$, then we have shown [3] that $L_{k,n}$ converges to R^*, the Bayesian error probability, almost surely.

Our main result is a bound on the asymptotic mean-square error $E(L_{k,n} - R^*)^2$ and a bound on $R_k - R^*$. These bounds on the error probability are not tight. For example, in the two-class case, the Cover–Hart bound [1] is tight and is much better than that presented here. It is not known how to extend their result to the case of multiple classes.

The Main Result

Let ξ be a random variable taking values in a separable metric space X with the metric d. Denote by ρ an integer-valued random variable taking values in $\{1, 2, \cdots, M\}$. The problem is to estimate ρ after observing ξ. The function $p_i(x) = P(\rho = i|\xi = x)$, $x \in X$, $i = 1, 2, \cdots, M$, is called the ith *a posteriori* probability function. Let $A_1, A_2, \cdots, A_M$ be the partition of the space X given by

$$A_i = \{x | p_i(x) \geq p_j(x), \text{ if } i \leq j, \; p_i(x) > p_j(x), \text{ if } i > j\},$$

Manuscript received March 15, 1976; revised October 28, 1977. This work was presented at the IEEE International Symposium on Information Theory, Ronneby, Sweden, 1976.

The authors are with the Technical University of Budapest, 1111 Budapest, Stoczek u. 2. Hungary.

0018-9448/78/0700-0512$00.75 ©1978 IEEE