INFORMATION AND CONTROL 16, 128–152 (1970)

A Generalized k-Nearest Neighbor Rule

E. A. PATRICK AND F. P. FISCHER, III

School of Electrical Engineering, Purdue University, Lafayette, Indiana 47907

A family of supervised, nonparametric decision rules, based on tolerance regions, is described which includes the k-Nearest Neighbor decision rules when there are two classes. There are two practical reasons for doing so: first, a family of decision rules similar to the k-Nearest Neighbor rules can be specified which applies to a broader collection of pattern recognition problems. This is because in the general class of rules constraints are weakened between the number of training samples required in each training sample set and the respective a priori class probabilities; and, a discrete loss function weighting the importance of the finite number of ways to make a decision error can be introduced.

Second, within the family of decision rules based on tolerance regions, there are decision rules which have a property allowing for preprocessing of the training set data resulting in significant data reduction.

Theoretical performance for a special case is presented.

I. INTRODUCTION

The general problem of supervised discrimination is considered for the following special case: A vector observation x is drawn from one of M classes denoted $\omega_1, \omega_2, \ldots, \omega_M$. If drawn from class ω_i, then x has a probability density function $f(x \mid \omega_i)$. P_i, the probability that x is drawn from class ω_i, is assumed known. $f(x \mid \omega_i)$ is assumed fixed, continuous, and unknown. From class ω_i, n_i independent training samples are assumed available: these samples are classified (i.e., the training or learning is called supervised). After receiving a sample of unknown class (which is not a member of a training group) called a candidate sample, the problem is to assign it to one of the M classes with minimum risk.

It is well-known that if the loss matrix with elements L_{ij} and underlying

statistics $f(x \mid \omega_i)$ are known,[1] the Bayes (minimum risk) decision rule chooses the class ω_k active if

$$\sum_{i=1}^{M} L_{ki} f(x \mid \omega_i) P_i = \min_{1 \leqslant j \leqslant M} \left\{ \sum_{i=1}^{M} L_{ji} f(x \mid \omega_i) P_i \right\}. \tag{1}$$

The objective of this section is to describe a general class of estimators of a p.d.f. based on the supervised samples and the use of these p.d.f.'s in a decision rule.

A generalization of the method of estimation proposed by Loftsgaarden and Quesenberry [1] will be used to estimate $f(x \mid \omega_i)$. Their estimate of a p.d.f. at a point x is as follows:

> Using n classified samples available from a single class p.d.f., find the distance r between x and the $v(n)$-th nearest neighbor of x. "Nearness" is measured by any convenient metric. Then

$$\hat{f}(x \mid \omega_i) = \frac{v(n) - 1}{n} \frac{1}{\Phi(x)}, \tag{2}$$

where $\Phi(x)$ is the volume of the set of all points whose distance to x is less than r. If

$$\lim_{n \to \infty} v(n) = \infty \qquad \text{and} \qquad \lim_{n \to \infty} v(n)/n = 0, \tag{3}$$

then they prove that $\hat{f}(x \mid \omega_i)$ approaches $f(x \mid \omega_i)$ with probability 1. Cover and Hart [2] suggested that the above estimator might be used in a decision rule to obtain a simple modification of the k-Nearest Neighbor (k-NN) decision rule. The k-NN rule, investigated by Fix and Hodges [3] for $k \to \infty$ and investigated by Cover and Hart [2] for fixed k, is a nonparametric procedure which assigns the candidate x^c the class which is most frequently represented in the k nearest neighbors to x^c. In Fig. 1, for example, the 5-th nearest neighbor decision rule would decide class ω_2 is active because four of the five nearest neighbors to x^c are from class ω_2.

The goal of this section is to describe a family of supervised, nonparametric decision rules, based on tolerance regions, which includes the previous k-NN decision rules when there are two pattern classes. There are two practical purposes for doing so: *First*, a family of decision rules similar to the k-NN rules can be specified which applies to a broader collection of pattern recognition problems. This is because, in the general class of rules, constraints are weakened between the number of training samples required

[1] For the problem considered herein, the underlying densities $f(x \mid \omega_i)$ are unknown.

in each training sample set and the respective a priori class probabilities; and, a discrete loss function weighting the importance of the finite number of ways to make a decision error can be introduced.

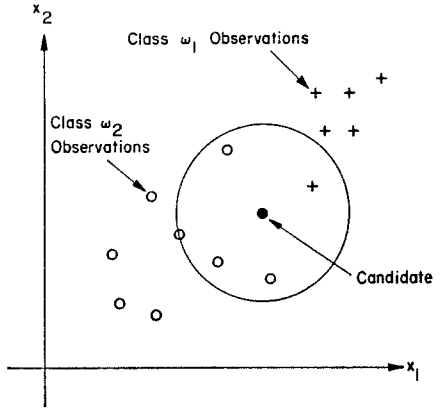

FIG. 1. Fifth Nearest Neighbor decision rule.

Second, within the family of decision rules based on tolerance regions, there are decision rules which have a property allowing for preprocessing of the training set data resulting in significant data reduction. Recognition based on such preprocessed data sets can be accomplished with a computer having limited storage capacity along with sequentially accessable memory, and the subtraction capability.

In the next section, the theory of distribution free tolerance regions is reviewed. In Section III, decision rules are described which use the properties of distribution free tolerance regions. The resulting decision rules and their implementation properties are described in Section IV, and theoretical performance for a special case is considered in Section V.

II. Distribution Free Tolerance Regions

The theory of statistical tolerance regions was initiated by Wilks [4]. Recent work in this area has been done by Kemperman [5], Fraser and Guttman [6], Fraser [7], and Wilks [8]. Fraser [7] provides an excellent discussion of tolerance regions in a general background. The following section is not to be construed as a survey. Although nothing new with

respect to the above papers is presented here, it permits readers unfamiliar with tolerance regions to become so.

In this section, we will be concerned with only a single p.d.f. $f(x)$, and a single set of training samples: let $x^1, x^2, \ldots, x^n$ be n independent, L-dimensional random observations with p.d.f. $f(x)$. Suppose we had a procedure which, for every possible way the n training samples could fall, gives a set of points in the observation space. A set $\mathscr{S}$ is called a tolerance region if it is a function of the training samples ([7], pp. 116).

For Example 1, one procedure might be: Let the set $\mathscr{S}$ equal the set of all points within an ellipsoid of concentration of the multivariate Gaussian distribution whose mean and covariance are the sample mean and covariance of the n observations.[2]

For Example 2, let $\mathscr{S}$ be the set of all points in the observation space which are nearer to x_0 than to any observation.

If $\mathscr{S}$ is an observed tolerance region, its coverage $P_{\mathscr{S}}$ is the probability that another sample drawn at random from $f(x)$ will fall in $\mathscr{S}$. That is,

$$P_{\mathscr{S}} = \int_{\mathscr{S}} f(x)\, dx.$$

Of course, $P_{\mathscr{S}}$ is bounded by 0 and 1. Because the n observations are drawn at random, the set $\mathscr{S}$ is a random set and, hence, its coverage $P_{\mathscr{S}}$ is a random real variable.

A tolerance region $\mathscr{S}$ is called a distribution-free tolerance region if the density function of $P_{\mathscr{S}}$ is independent of the underlying p.d.f. $f(x)$. It turns out as might be expected, that in the first example, above, the tolerance region is distribution free within the family of Gaussian p.d.f.'s. One would expect that distribution-free tolerance regions would be difficult to construct. Due to Wilks and authors who followed, this is seen not to be the case.

Next we give one simple form of Tukey's construction [9]. With it, we can show that the tolerance region of Example 2 is distribution free. The following rule will be used when we receive n L-dimensional observations from unknown p.d.f. $f(x)$. Prior to knowledge of observations, an arbitrary real valued, noninfinite, continuous function $\phi(x)$ defined on the observation space, and a positive integer less than $n + 1$ is specified. When the n

[2] Knowledge of ellipsoid shape for local regions in the observation space is very important. Effectively, the use of a local ellipsoid corresponds to local dimensionality reduction. Success in pattern recognition may depend on how much is known a priori about the local ellipsoid shape.

observations are received, we will evaluate $\phi(x)$ at each of the observation points. Associated with an observation x^j is its "order" $\phi(x^j)$. The observations are ranked according to their "order" from smallest to largest. Letting z equal the v-th smallest "order value", the set of all the points in the observation space whose "order value" is less than z is defined to be the tolerance region $\mathscr{I}$.

It results that the coverage $P_{\mathscr{I}}$ for this construction always has the Beta distribution $Be(v, n - v + 1)$, independent of the underlying p.d.f. $f(x)$. If X has the $Be(v, n - v + 1)$ distribution, X has the density

$$f(x) = \begin{cases} \dfrac{n!}{(v-1)!\,(n-v)!}\, x^{v-1}(1-x)^{n-v} & 0 \leqslant x \leqslant 1 \\ 0 & \text{elsewhere} \end{cases} \tag{5}$$

To show that Example 2 is a distribution-free tolerance region, let $\phi(x)$ equal the Euclidian distance from a fixed point x_0 to x, and let v equal some fixed positive integer less or equal to n. The rule will be: $\mathscr{I}$ will be the set of all points inside the hypersphere centered at x_0 which contains $v - 1$ observations inside, one observation on the surface (which is not in $\mathscr{I}$), and $n - v$ observations outside. This is the scheme used in [1]. A simple modification is to define $\phi(x) = x^T\Sigma^{-1}x$, where Σ is a positive definite symmetric matrix. Then $\mathscr{I}$ would be a hyperellipse.

Just as one tolerance region can be formed, the whole observation space can be partitioned into s nonoverlapping tolerance regions [9]. All that is required is that a list of s functions and integers be specified. Essentially, the same procedure is followed, except that tolerance regions at later stages of the construction must lie totally within tolerance regions of initial stages.

For Example 3, suppose we have the following simple rule for six 2-dimensional observations (see [9], p. 529): Order the observations by their first coordinates [$\phi(x) = x_1$, where $x = (x_1, x_2)$]. Let h_1 be the value of the smallest x_1 value ($v_1 = 1$). The first tolerance region is the set of all points x such that $x_1 \leqslant h_1$. This is just the region to the left of the leftmost observation. Now, the second order function $\phi_2(x) = x_2$ is used to order the remaining observations by their value x_2. The first smallest value of x_2 is denoted h_2, and the second tolerance region is the set of points $x = (x_1, x_2)$ such that $x_1 < h_1$, and $x_2 < h_2$. This procedure can be continued to remove regions counter clockwise about the space, using $\phi(x) - x_1$, or $-x_2$ on the other sides. The resulting partition of the observation space is shown in Fig. 2.

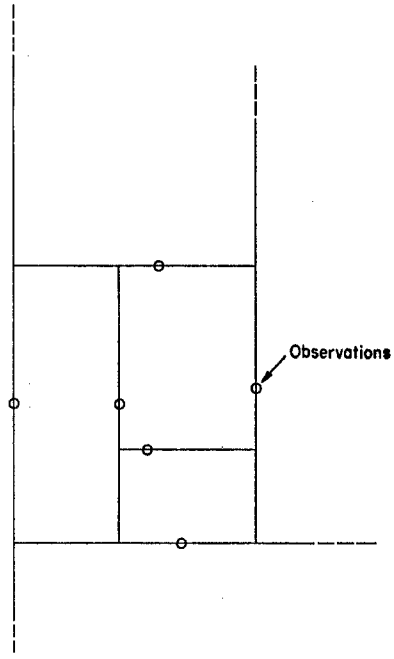

FIG. 2. Partitioning of observation space.

It should be pointed out that the use of ordering functions in Example 3 is not a consistent procedure. That is, as the number of available training samples increases, the longest diagonal of most tolerance regions does not go to zero. Hence, in the limit,

$$\max_{x,y\in\mathscr{S}_i} |f(x) - f(y)| \neq 0.$$

This property is required of $\mathscr{S}_i$ in the next section. This problem is easily alleviated by subpartitioning tolerance regions to obtain the above property. One such procedure (Example 4) is to partition the space into $L + 1$ regions where each region contains approximately the same number of observations by using a modification of Example 3. Then, each tolerance region is processed in a similar manner with a cyclical use of the order functions until each region contains v samples. Because this procedure generates many infinite volume regions, the construction is started after bounding the samples by $2L$ hyperplanes. An illustration is provided in Fig. 3 for $v = 1$, $n = 16$.

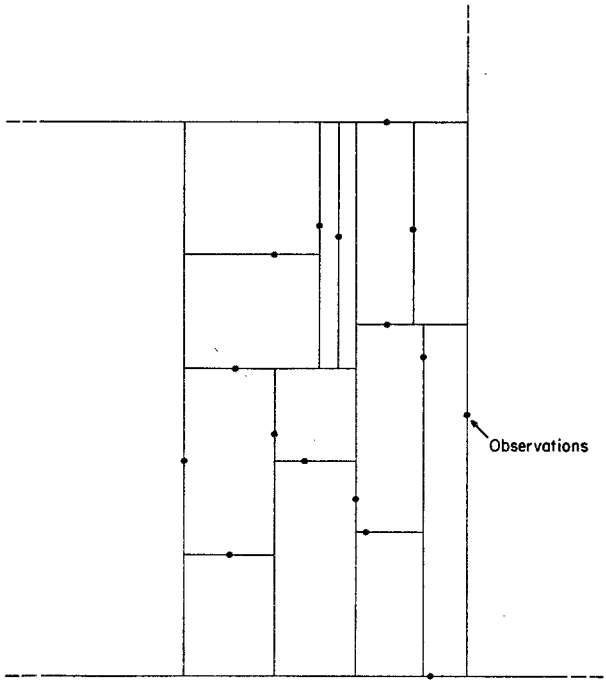

FIG. 3. Subpartitioning of observation space.

III. ESTIMATION

As before, it is now assumed that there are classified sets of training samples from each of M classes; a candidate observation, x^c, of unknown class is available. The problem is to assign x^c to one of the M classes.

The approach using partitions formed by constructing sequential tolerance regions discussed in the last section is as follows: The training samples from the i-th class are processed to form tolerance regions which partition the observation space. The location of the tolerance regions and the statistics of the coverages of each region is the information which is learned about the i-th underlying class conditional density function. This information is used to form an estimate probability density at each point x of the observation space. To simplify notation, let ξ_i be the index of the tolerance region for the i-th class which contains x. If x happens to be on a boundary between tolerance regions, we arbitrarily will choose the smaller index. (Since the

volume of all boundaries is zero and hence contains zero probability, this choice will be made with zero probability if X is a random variable having a density with no discrete component.)

It is known a priori that the coverage of the ξ_i-th tolerance region has the Beta distribution $Be(v_{\xi_i}, n_i - v_{\xi_i} + 1)$, where v_{ξ_i} is the number of samples involved in constructing a tolerance region using Tukey's construction. We know that the average coverage of tolerance region ξ_i is the mean of the Beta distribution, or $v_{\xi_i}/(n_i + 1)$. Thus, we define the estimate of the density at x to be the ratio of the expected coverage of the tolerance region containing x to the volume Φ_{ξ_i} of the tolerance region:

$$\hat{f}(x \mid \omega_i) = \left(\frac{v_{\xi_i}}{n_i + 1}\right) \Big/ \Phi_{\xi_i}. \tag{6}$$

The above estimate is asymptotically identical with (2). Loftsgaarden and Quesenberry proved that (2) is an asymptotically unbiased estimate provided that the maximum diameter of the tolerance regions converges to 0. Thus, (6) also is asymptotically unbiased if x is a continuity point of f and the maximum diameter of the tolerance regions converges to zero. The difference in performance is evident only with finite sample size.

IV. Decision Rules and Implementation

The decision rule is obtained by replacing $f(x \mid \omega_i)$ in Eq. (1) by $\hat{f}(x \mid \omega_i)$ [Eq. (6)], resulting in:

Choose class ω_k active if

$$\sum_{i=1}^{M} L_{ki} \hat{f}(x \mid \omega_i) P_i = \min_{1 \leq j \leq M} \sum_{i=1}^{M} L_{ji} \hat{f}(x \mid \omega_i) P_i, \tag{7}$$

which by (6) is equivalent to

$$\sum_{i=1}^{M} L_{ki} \frac{v_{\xi_i}}{n_i + 1} \frac{P_i}{\Phi_{\xi_i}} = \min_{1 \leq j \leq M} \sum_{i=1}^{M} L_{ji} \frac{v_{\xi_i}}{n_i + 1} \frac{P_i}{\Phi_{\xi_i}}. \tag{8}$$

For the special case when the loss function weights all types of errors equally, and a correct decision has zero loss, Eq. (7) reduces to:

Choose class ω_k active if

$$\hat{f}(x \mid \omega_k) P_k = \max_{1 \leq j \leq M} \hat{f}(x \mid \omega_j) P_j, \tag{9}$$

which is equivalent to

$$\frac{v_{\xi_k} P_k}{(n_k + 1)\,\Phi_{\xi_k}} = \max_{1 \leqslant j \leqslant M} \frac{v_{\xi_j} P_j}{(n_j + 1)\,\Phi_{\xi_j}}. \tag{10}$$

Patrick [10] shows that the decision rule (8) or (10) minimizes the risk conditioned on the location of tolerance regions and the unclassified candidate sample.[3] An outline of a proof is provided in Appendix I. We now will point out interesting properties of two special cases of the above rules.

Case 1. Suppose that the spherical tolerance regions of Example 2 are used in (10) for the case $M = 2$ and $P_1/P_2 = (n_1 + 1)/(n_2 + 1)$. This is the binary, $0 - 1$ loss function case when the number of training samples from each class are representative of the a priori class probabilities. Then (10) reduces to:

Choose class ω_k active if

$$\frac{v_{\xi_k}}{\Phi_{\xi_k}} = \max_{j=1,2} \frac{v_{\xi_j}}{\Phi_{\xi_j}}. \tag{11}$$

Also, suppose, for the moment, that the same tolerance region construction is used for both classes so that $v_{\xi_1} = v_{\xi_2} = v$. Then, the volume Φ_{ξ_1} will be smaller than the volume of Φ_{ξ_2} if, and only if, the v-th class 1 sample is closer to x^c than the v-th class 2 sample is to x^c. But this is equivalent to choosing the class which is most highly represented among the first $2v + 1$ nearest neighbors of the pooled samples. This is also equivalent to choosing the class which is most highly represented among the first $2v$ nearest neighbors, providing that a tie does not occur. Thus, this special case of decision rules based on tolerance regions gives exactly the same decision as the k-Nearest Neighbor rule, where $v = [(k + 1)/2]^+$, with the exception mentioned for the case when k is even. Rule (10) is more general than the k-NN rule in the sense that weights can be attributed to different types of errors, and problems can be handled in which the training samples available are not in the same proportions as the a priori class probabilities. Further, different circular tolerance region construction methods may be used for each class. This is provided by allowing for the inequality of v_{ξ_1} and v_{ξ_2}.

An example of a problem in which the k-NN rule cannot be applied is as follows.

[3] The unclassified candidate sample is the sample to be recognized using the generalized k — NN decision rule. In this paper, it is sometimes denoted as x^c.

Suppose we know that $P_1 = P_2 = 1/2$. However, the training samples available for the two classes are of size $n_1 = r$ and $n_2 = 3r$. In this case, if the k-NN rule were applied, the algorithm would "learn" that $P_1 = 1/4$ and $P_2 = 3/4$ with a resultant degradation in performance. The generalized k-NN rule using circular tolerance regions would easily take this into account.

The generalized k-NN rule offers more experimental freedom than the k-NN decision rule because, in the former, the value of v_{ξ_i} may be different for each class.

Case 2. We now discuss a second special case of decision rules (10) based on tolerance regions. Suppose that once the training samples are received, the observation space is partitioned for each class. The location of the regions will be independent of the candidate. The result is that, instead of creating tolerance regions after each candidate, as in Case 2, the process need be completed only once (i.e., presuming that the training set has not changed). We are going to consider construction techniques such as illustrated in Example 3 which lend themselves to this procedure.

In Example 3 of Section II, the sequence of numbers $h_1, h_2, \ldots, h_{s_i}$ determined the location of all of the tolerance regions in a simple manner for a single class. Thus, the sequence can be thought of as "coding" or representing the estimate density $\hat{f}(x)$, with the understanding that the rule for constructing the tolerance regions is known.

The function $\hat{f}(x \mid \omega_i)$ is evaluated at x, using the preprocessed data by first locating the tolerance region in which x lies. This is done by making a sequence of at most s_i differences of real numbers. The tolerance region volume is determined by the tolerance region's boundaries.

Because the method of constructing tolerance regions is assumed known at recognition time, the v_{ξ_i} are known for all i and x. Thus, all of these parameters are obtainable from the reduced data set for substitution into Eq. (10). As a consequence of storing the tolerance region parameters rather than the original data, data reduction-results.

An alternative procedure is to generate for each tolerance region of each class, during preprocessing, a sequence whose elements are a monotonic function of $P_i v_{\xi_i}/(n_i + 1)$. Recognition would then consist of finding the index of the tolerance regions containing x in each class and then choosing the class with the highest $P_i v_{\xi_i}/(n_i + 1)$ which are found in the table. One such monotonic function would be the natural order of the $s = \sum_{i=1}^{M} s_i$ numbers. This particular list could be stored in a condensed list.

The practical advantages of preprocessing the data to obtain tolerance region parameters are multifold: The primary advantage is that of complexity

reduction in implementing the decision rule. Instead of calculating $\hat{f}(x \mid \omega_i)$ using a Case 1 generalized k-NN rule based on n_i, L-dimensional observations per class, which consumes copious amounts of computer storage and computation time, the Case 2 generalized k-NN rule is based on only s_i, 1-dimensional observations per class where $s_i = n_i$. Thus, preprocessing to obtain tolerance region boundaries reduces the amount of storage required to implement the decision rule by a factor of at least $1/L$. If $v_{\xi_i} = k$ for all i and x, reduction by a factor of $1/kL$ takes place ($2/kL$ for the alternative procedure). Thus, large storage reductions can be obtained at the expense of preprocessing.

As a result of data reduction, the complexity of calculation at recognition time is greatly reduced. In addition to the fact there are fewer numbers to process, the computations involve only subtraction if a preprocessed volume list is available under the alternative procedure.

It is possible that a volume Φ_i in the density estimate (6) or the decision rule (8) will be infinite. In such a case, $\hat{f}(x \mid \omega_i)$ is zero.

V. Theoretical Performance

The objective is to *outline* a calculation of theoretical performance of the generalized k-NN decision rule using circular tolerance regions[4] against any two underlying distributions (the performance will clearly be a function of the two distributions. This objective is accomplished by first calculating the probability of error ($0 - 1$ loss function) incurred through use of the decision rule at every point in the observation space. The performance is then the average of the point performance with respect to the mixture distribution (only point performance is considered in this paper).

The problem is described by the following experiment: suppose an experimenter has at his disposal a procedure to draw random vector samples from either of two distributions, and x^c (x^{n+1}, the sample being tested, will henceforth be called the candidate sample x^c) is a point fixed in the observation space. The underlying continuous c.d.f.'s and specifically, $W_1 = f(x^c \mid \omega_1)$, $W_2 = f(x^c \mid \omega_2)$, are known to the experimenter.

With no loss of generality, assume $P_1W_1 > P_2W_2$. The experimenter will draw n_i supervised random vector samples from each of the two underlying distributions, thereafter presenting them as training samples characterizing the underlying distribution function (unknown to the GK estimation system). Clearly, the experimenter knows that the minimum probability of error rule will choose class ω_1, because $P_1W_1 > P_2W_2$.

[4] For convenience we will refer to this rule as the GK rule.

Let $P_{2,1}$ be the probability the GK decision rule choose class ω_2, when in fact class ω_1 is most probable. P_e, the probability of error if x^c were caused by class ω_1 with probability $P_1W_1/(P_1W_1 + P_2W_2) = \eta_1$ can then be calculated in terms of $P_{2,1}$.

Now, we analyze what happens at recognition time. The GK system accepts n_i supervised vector samples and obtains the location of the tolerance regions partitioning the observation space for class ω_i, $i = 1, 2$. By definition, the indexes of the two tolerance regions $\mathscr{I}_{\xi_1}$ and $\mathscr{I}_{\xi_2}$ containing x^c are always the same: $\xi_1 = \xi_2 = 1$. Let U_i and Φ_i be, respectively, the coverage of $\mathscr{I}_{\xi_i}$ with respect to $f(x \mid \omega_i)$ and the volume of $\mathscr{I}_{\xi_i}$, $i = 1, 2$. Then, the GK decision [according to Eq. (10)] will be:

Choose class ω_2 if

$$\frac{P_1v_1}{(n_1 + 1)\,\Phi_1} < \frac{P_2v_2}{(n_2 + 1)\,\Phi_2}.$$

We will calculate the conditional probability of the above event, conditioned on P_1W_1, P_2W_2, at point x^c, where

$$W_1 = f(x^c \mid \omega_1)$$
$$W_2 = f(x^c \mid \omega_2).$$

The calculation will be made under the assumption that $P_1W_1 > P_2W_2$. Denote this conditional probability $P_{2,1}$:

$$P_{2,1} = P\left[\frac{P_1v_1}{(n_1 + 1)\,\Phi_1} < \frac{P_2v_2}{(n_2 + 1)\,\Phi_2}\right].$$

To determine this probability, the distribution law on Φ_1 and Φ_2 must be determined. To accomplish this, an assumption is made that the tolerance regions are "sufficiently small" such that it is reasonable to state: $U_i = W_i\Phi_i$. This approximation will be poor for *small training sample sets* for most underlying distributions.[5]

Hence,

$$P_{2,1} \approx P\left[\frac{n_1 + 1}{P_1v_1W_1}\,U_1 > \frac{n_2 + 1}{P_2v_2W_2}\,U_2\right].$$

[5] Note that the tolerance region can only be "sufficiently small" with some hopefully high probability. It appears, however, that a theoretical performance result can be obtained without the uniformity assumption.

The probability of this event is calculable since U_1 and U_2 have the beta distribution, independent of the underlying p.d.f.'s. Define

$$R_i = U_i/\theta_i = \frac{n_i + 1}{P_i v_i W_i} U_i, \qquad i = 1, 2.$$

Then R_i is a random variable having the scaled beta distribution:

$$f_i(u) = \frac{n_i!}{(v_i - 1)!\,(n_i - v_i)!}\,\theta_i(u\theta_i)^{v_i-1}(1 - u\theta_i)^{n_i-v_i}; \quad 0 < u < 1/\theta_i, \ \theta_i > 0$$

and

$$P_{2,1} \cong P[R_1 > R_2] = \begin{cases} \int_0^{1/\theta_1} \left\{ \int_0^s f_1(s) f_2(t)\, dt \right\} ds & \text{if} \quad \theta_1 > \theta_2 \\ 1 - \int_0^{1/\theta_2} \left\{ \int_0^t f_1(s) f_2(t)\, ds \right\} dt & \text{if} \quad \theta_1 < \theta_2 . \end{cases} \tag{11}$$

The above integrals are evaluated in Appendix II with the result[6]

$$P_{2,1} \cong \begin{cases} \dfrac{n_1!\,n_2!}{(v_1 - 1)!\,(v_2 - 1)!} \displaystyle\sum_{j=0}^{n_2-v_2} \frac{(-1)^j \left(\dfrac{\theta_2}{\theta_1}\right)^{v_2+j} (v_2 + v_1 + j - 1)!}{j!(n_2 - v_2 - j)!\,(v_2 + j)(v_2 + n_1 + j)!} & \text{if} \quad \theta_1 > \theta_2 \\ 1 - \dfrac{n_1!\,n_2!}{(v_1 - 1)!\,(v_2 - 1)!} \displaystyle\sum_{j=0}^{n_1-v_1} \frac{(-1)^j \left(\dfrac{\theta_1}{\theta_2}\right)^{v_1+j} (v_1 + v_2 + j - 1)!}{j!(n_1 - v_1 - j)!\,(v_1 + j)(v_1 + n_2 + j)!} & \text{if} \quad \theta_1 < \theta_2 \end{cases} \tag{12}$$

For the special case when $v_1 = v_2 = v$ and $n_1 = n_2 = n$, and n approaches infinity, the limiting form of $P_{2,1}$ (Appendix III) has been found:

$$P_{2,1} = \frac{1}{\left(1 + \dfrac{P_2 W_2}{P_1 W_1}\right)^{2v-1}} \sum_{j=v}^{2v-1} \binom{2v - 1}{v} \left(\frac{P_2 W_2}{P_1 W_1}\right)^j; \qquad \begin{matrix} n_1 = n_2 = \infty \\ v_1 = v_2 = v. \end{matrix} \tag{13}$$

[6] Note that (11) and (12) presume the assumption $U_i = W_i \Phi_i$. Except where $W_i = 0$, this assumption precludes the possibility of infinite volume Φ_i.

Since $P_{2,1}$ is the probability that the decision rule will choose class 2 evaluated under the assumption that class 1 is more probable, and $P_2W_2/(P_1W_1 + P_2W_2) = \eta_2$ is the probability the vector x^c actually is from class ω_2, the probability of error (or misclassification of x^c) for this experiment is found by substituting (13) into the equation below:

$$P_e = \eta_2(1 - P_{2,1}) + (1 - \eta_2)(P_{2,1}). \tag{14}$$

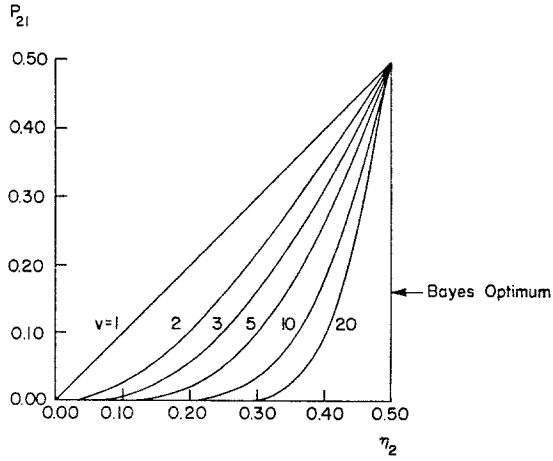

FIG. 4. Asymptotic Bayes-Conditional Error. ($n = \infty$)

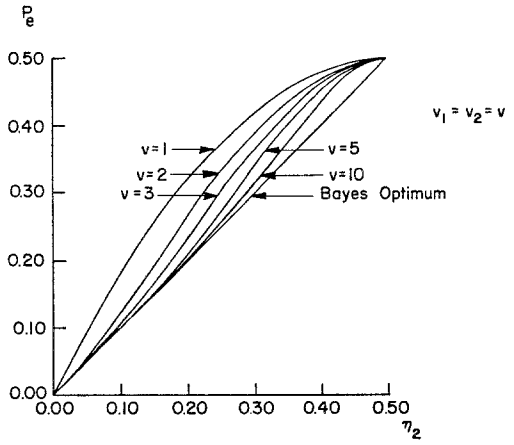

FIG. 5. Asymptotic probability of error. ($n = \infty$)

The probability $P_{2,1}$ of choosing the less likely class and P_e, the probability of error for the asymptotic case, are plotted in Figs. 4 and 5 against the probability of error if the statistics had been known.

It is easy to show that for the special case of $v_1 = v_2 = v$, $P_1 = P_2$, the generalized k-NN decision rule has the same risk as the k-NN indicated by Cover and Hart [2] for $k = 2v - 1$. This might have been guessed because the generalized k-NN decision rule using circular tolerance regions gives the same decision as the k-NN rule for the above mentioned special case.

Theoretical small sample performance is not distribution free. Small sample performance is an open research problem. Experimental small sample performance may be found in Ref. [11]. The result (14) for probability of error is at the point x^c; an extension would be to obtain probability of error by averaging the point performance on the mixture distribution.

VI. Example

An example illustrating application of the generalized k-Nearest Neighbor decision rule is illustrated in Fig. 6 for a two class case. Let $X^1, X^2, \ldots, X^{n_1}$ be n_1 training samples from class 1 and $Y^1, Y^2, \ldots, Y^{n_2}$ be n_2 training samples from class 2. A priori supply the class probabilities P_1, P_2, and metrics $\rho(x, x^c), \rho(y, x^c)$ for the respective classes. Given that there are to be v_1 class 1 samples and v_2 class 2 samples in two respective tolerance regions centered at x^c, compute the tolerance region volumes Φ_1 and Φ_2. For example, Φ_1 is computed after finding the v_1 nearest samples to x^c using the metric $\rho(x, x^c)$.

Experimentally, it has been found very desirable to shape the metrics ρ to fit the data's covariance structure. Because the data's covariance structure can vary for different points x^c, it may be desirable to partition the observation space into regions and assign local metrics to the respective regions.[7] (One way to partition uses the straight line ordering functions described in Section II.)

Another modification of the system shown in Fig. 6 would be to incorporate a provision for estimating or adapting the metrics with a priori starting

[7] However, the reader should be warned that this double-use of the data invalidates the theorem. Incidently, the problem is very nearly similar to the parametric estimation problem of estimating the gaussian component density in a mixture using unsupervised estimation [12].

metrics. *We conjecture that a priori starting metrics reflect a priori knowledge about the problem model and are very important for solving pattern recognition problems.*

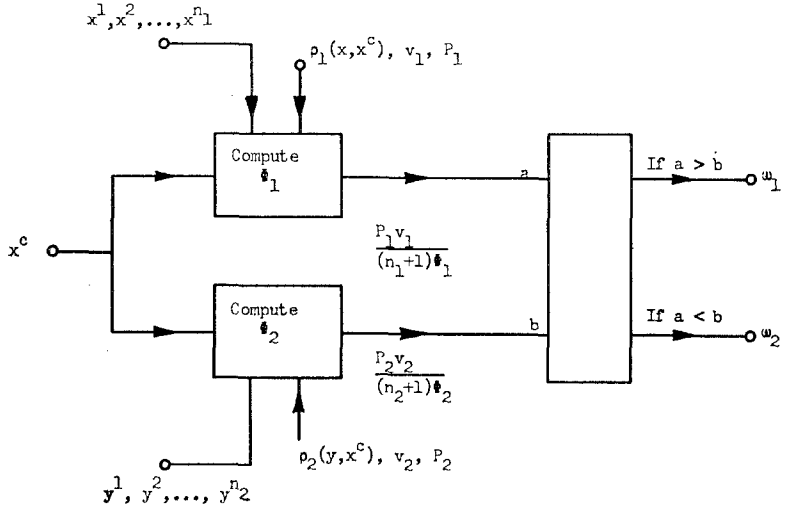

FIG. 6. Example using circular tolerance regions.

It has been observed for a typical problem, where $f(x \mid \omega_i)$ may be multimodal, $L = 6$, and $n_1 = n_2 = 500$, that experimental performance depends critically on v_1 and v_2. Reasonable values might be $v_1 = v_2 = 3$. An explanation is that for small sample sizes, if $v_1 = v_2 = 1$, there are too few samples for local density estimation; and if $v_1 = v_2 = 15$, the local density concept does not apply.

CONCLUSIONS

A generalized k-NN decision rule is presented which utilizes local density estimation (2) in the decision rule (1). It may be appropriate to call this generalized k-NN decision rule the k-NN_3 rule to distinguish it from the decision rule studied by Cover and Hart [2] and the decision rule studies by Fix and Hodges [3]. Then, the decision rule studied by Cover and Hart could be called the k-NN_2 decision rule, and the one studied by Fix and Hodges the k-NN_1 decision rule.

Properties of the generalized k-NN decision rule (k-NN_3 rule) are discussed in the paper.

APPENDIX I. Outline of Minimum Conditional Risk Argument

Let η be the class of all rules for obtaining s_i *distribution-free tolerance regions*[8] in the L-dimensional observation space I^L, given n_i samples $x^1,\ldots, x^{n_i}$ from class ω_i. Denote the ξ-th tolerance region by $\mathscr{I}_{\xi_i}$; then

$$I^L = \bigcup_{\xi=1}^{s_i} \mathscr{I}_{\xi_i} n = \sum_{i=1}^{M} n_i \tag{1}$$

because distribution-free tolerance regions cover the L-dimensional space and they are mutually exclusive. In like manner, the observation space for each of the M classes is covered by distribution-free tolerance regions, such that

$$s = \sum_{i=1}^{M} s_i , \tag{2}$$

where s is the total number of distribution-free tolerance regions, formed using n supervised vector samples.

In order to keep the notation as simple as possible, we will not provide for indexing the unlimited number of ways or rules for forming distribution free tolerance regions. We will, however, distinguish between the rules used for the respective classes; thus, let η_i be the rule for obtaining distribution-free tolerance regions for the i-th class. Let η denote the M rules $\eta_1, \eta_2, \ldots, \eta_M$, $\eta_i \in \underline{\eta}$.

Assume that I^L consists of elementary, disjoint regions whose union is I^L. A specific elementary region is denoted I_t with corresponding volume φ. The event that observation x^s is in I_t is denoted x_t^s. Denote by Y_t^n the sequence $(x_t^1, x_t^2,\ldots, x_t^n)$, where I_t for x_t^1 is not necessarily the same as I_t for x_t^2, etc.

The problem is to decide, with minimum risk, which class caused x_t^{n+1} given supervised samples Y_t^n and η. By a straightforward extension of (7),

[8] Properties of distribution-free tolerance regions are discussed in Section II. A distribution-free tolerance limit should not be confused with a distribution-free tolerance region; the former is a statement concerning the probability mass (or coverage) in the distribution free tolerance region.

risk conditioned on Y_t^n and η is minimized by the following decision rule: Decide ω_a active if

$$\sum_{i=1}^{M} L_{ai} p(\omega_i, x_t^{n+1} \mid Y_t^n, \eta_i) = \min_K \left\{ \sum_{i=1}^{M} L_{Ki} p(\omega_i, x_t^{n+1} \mid Y_t^n, \eta_i) \right\}_{K=1}^{M}. \quad (3)$$

The term $p(\omega_i, x_t^{n+1} \mid Y_t^n, \eta_i)$ can be evaluated according to the following lemma.

LEMMA 1. *If n_i supervised samples form s_i distribution-free tolerance regions using rule η_i, and sample x^{n+1} is in tolerance region $\mathscr{I}_{\xi_i}$, having volume Φ_{ξ_i} and coverage $\mathscr{P}_{\xi_i}$, then*

$$p(\omega_i, x_t^{n+1} \mid Y_t^n, \eta_i) = \frac{\varphi}{\Phi_{\xi_i}} E[P_i \mathscr{P}_{\xi_i} \mid Y_t^n, \eta_i] \quad (4)$$

or if $\delta_n{}^i$ is the set of tolerance regions[9] for class ω_i $[\delta_n{}^i = (Y_t^n, \eta_i)]$

$$p(\omega_i, x_t^{n+1} \mid \delta_n{}^i) = \frac{\varphi}{\Phi_{\xi_i}} E[P_i \mathscr{P}_{\xi_i} \mid \delta_n{}^i]. \quad (5)$$

Proof. The lemma is proven in two parts. We will denote the event $x_t^{n+1} \in \mathscr{I}_{\xi_i}$ by $\mathscr{R}_{n+1\xi_i}$, and let $\mathscr{L}_n{}^i$ denote the sequence of n_i tolerance regions for class ω_i successively containing samples $x^1, x^2, \ldots, x^{n_i}$. The first part of the proof of Lemma 1 follows.

PART 1.

$$p(\omega_i, x_t^{n+1} \mid \mathscr{L}_n^i, \eta_i) = \frac{\varphi}{\Phi_{\xi_i}} p(\omega_i, \mathscr{R}_{n+1\xi_i} \mid \mathscr{L}_n^i, \eta_i) \quad (6)$$

for all $x_t^{n+1} \in \mathscr{I}_{\xi_i}$.

Proof. If the event (ω_i, x_t^{n+1}) occurs, then so does the event $\mathscr{R}_{n+1\xi_i}$ because $x_t^{n+1} \subset \mathscr{R}_{n+1\xi_i}$; in other words, the event $(\omega_i, x_t^{n+1}, \mathscr{R}_{n+1\xi_i})$ occurs. Thus,

$$\begin{aligned} p(\omega_i, x_t^{n+1} \mid \mathscr{L}_n^i, \eta_i) &= p(\omega_i, x_t^{n+1}, \mathscr{R}_{n+1\xi_i} \mid \mathscr{L}_n^i, \eta_i) \\ &= p(x_t^{n+1} \mid \mathscr{R}_{n+1\xi_i}, \omega_i, \mathscr{L}_n^i, \eta_i)\, p(\mathscr{R}_{n+1\xi_i}, \omega_i \mid \mathscr{L}_n^i, \eta_i). \end{aligned} \quad (7)$$

[9] The set of tolerance regions (the partition denoted $\delta_n{}^i$) is uniquely determined by the n training samples Y_t^n and the tule η_i for forming the tolerance regions.

Now consider the term $p(x_t^{n+1} \mid \mathscr{R}_{n+1\xi_i}, \omega_i, \mathscr{L}_n^i, \eta_i)$. $\mathscr{L}_n^i$ adds nothing to estimating conditional probability mass in I_t, given $\mathscr{I}_{\xi_i}$, because the event $\mathscr{I}_{\xi_i}$ is a complexity constraint prohibiting fine knowledge of the probability structure within $\mathscr{I}_{\xi_i}$. Furthermore, we have no reason to favor any region $I_t \in \mathscr{I}_{\xi_i}$ with more mass than any other. Thus, each elementary region in $\mathscr{I}_{\xi_i}$ has equal mass with the upshot that

$$p(x_t^{n+1} \mid \mathscr{R}_{n+1\xi_i}, \omega_i, \mathscr{L}_n^i, \eta_i) = p(x_t^{n+1} \mid \mathscr{I}_{\xi_i}, \omega_i) = \frac{\varphi}{\Phi_{\xi_i}} \tag{8}$$

since $(\mathscr{R}_{n+1\xi_i}, \eta_i) = (\mathscr{I}_{\xi_i})$. This concludes Part 1.

Next we show that $p(\mathscr{R}_{n+1\xi_i}, \omega_i \mid \mathscr{L}_n^i, \eta_i)$, in relation (7), is simply the conditional expectation of the product $P_i\mathscr{P}_{\xi_i}$, where $\mathscr{P}_{\xi_i}$ is the probability mass (coverage) in tolerance region $\mathscr{I}_{\xi_i}$.

PART 2.

$$p(\mathscr{R}_{n+1\xi_i}, \omega_i \mid \mathscr{L}_n^i, \eta_i) = E[P_i\mathscr{P}_{\xi_i} \mid \mathscr{L}_n^i, \eta_i], \tag{9}$$

where

$$\mathscr{P}_{\xi_i} = p[x_t^{n+1} \in \mathscr{I}_{\xi_i} \mid \omega_i] = p[\mathscr{R}_{n+1\xi_i} \mid \omega_i].$$

Proof. Since $p(\mathscr{R}_{n+1\xi_i}, \omega_i)$ is completely characterized by $\mathscr{P}_{\xi_i}P_i$, it follows that

$$p(\mathscr{R}_{n+1\xi_i}, \omega_i \mid \mathscr{L}_n^i, \eta_i) = \int P_i\mathscr{P}_{\xi_i}\, dF(P_i, \mathscr{P}_{\xi_i} \mid \mathscr{L}_n^i, \eta_i)$$

$$= E[P_i\mathscr{P}_{\xi_i} \mid \mathscr{L}_n^i, \eta_i].$$

Relations (8) and (9) inserted in (7) give the desired result,

$$p(\omega_i, x_{n+1_t} \mid \mathscr{L}_n^i, \eta_i) = \frac{\varphi}{\Phi_{\xi_i}} E[P_i\mathscr{P}_{\xi_i} \mid \mathscr{L}_n^i, \eta_i].$$

This concludes the proof of Lemma 1.

Inserting (4) in (3) gives: Decide ω_a active if

$$\sum_{i=1}^{M} \frac{L_{ai}}{\Phi_{\xi_i}} E[P_i\mathscr{P}_{\xi_i} \mid Y_t^n, \eta_i] = \min_K \left\{\sum_{i=1}^{M} \frac{L_{Ki}}{\Phi_{\xi_i}} E[P_i\mathscr{P}_{\xi_i} \mid Y_t^n, \eta_i]\right\}_{K=1}^{M}; \tag{10}$$

Or equivalently: Decide ω_a active if

$$\sum_{i=1}^{M} \frac{L_{ai}}{\Phi_{\xi_i}} E[P_i \mathscr{P}_{\xi_i} \mid \delta_n{}^i] = \min_K \left\{ \sum_{i=1}^{M} \frac{L_{Ki}}{\Phi_{\xi_i}} E[P_i \mathscr{P}_{\xi_i} \mid \delta_n{}^i] \right\}_{K=1}^{M}. \tag{11}$$

If $\mathscr{P}_{\xi_i}$ is the coverage of a tolerance region formed using v_{ξ_i} samples, where the total number of samples is n_i, then the expected value of the coverage is $v_{\xi_i}/(n_i + 1)$. In terms of our notation,

$$E[\mathscr{P}_{\xi_i} \mid Y_i{}^n, \eta_i] = \frac{v_{\xi_i}}{n_i + 1}. \tag{12}$$

Inserting (12) in (10) or (11) and assuming P_i known, the resulting decision rule is: Decide ω_a active if

$$\sum_{i=1}^{M} P_i \frac{L_{ai}}{\Phi_{\xi_i}} \frac{v_{\xi_i}}{n_i + 1} = \min_K \left\{ \sum_{i=1}^{M} P_i \frac{L_{Ki}}{\Phi_{\xi_i}} \frac{v_{\xi_i}}{n_i + 1} \right\}_{K=1}^{M}. \tag{13}$$

And if $L_{ji} = 1$ for $j \neq i$ and zero, otherwise, relation (13) reduces to decide ω_a active if

$$\frac{P_a}{\Phi_{\xi_a}} \frac{v_{\xi_a}}{n_a + 1} = \max_K \left\{ \frac{P_K}{\Phi_{\xi_K}} \frac{v_{\xi_K}}{n_K + 1} \right\}_{K=1}^{M}. \tag{14}$$

The results of Lemma 1 through Eq. (14) can be summarized according to the following theorem.

THEOREM 1. *If n_i supervised samples form s_i distribution-free tolerance regions using rule η_i, then x^{n+1} is in tolerance region $\mathscr{I}_{\xi_i}$, having volume Φ_{ξ_i} and coverage $\mathscr{P}_{\xi_i}$, and if $p(x \mid \omega_i)$ is uniform in $\mathscr{I}_{\xi_i}$, then risk conditioned on Y^n and rule η is minimized by using decision rule* (14).

The set η, against which risk is minimized, has not been shown to be optimum. However, the use of η (which replaces the samples with distribution-free tolerance regions) appears to be an engineering approach having practical application and merits further consideration.

APPENDIX II

We wish to evaluate the integral

$$I = \begin{cases} \displaystyle\int_0^{1/\theta_1} \int_0^{s} f_1(s) f_2(t)\, dt\, ds : \theta_1 > \theta_2 \\ \displaystyle 1 - \int_0^{1/\theta_2} \int_0^{t} f_1(t) f_2(s)\, ds\, dt : \theta_1 < \theta_2 \end{cases}$$

where

$$f_i(u) = \frac{n_i!}{(v_i-1)!\,(n_i-v_i)!}\,\theta_i(u\theta_i)^{v_i-1}(1-u\theta_i)^{n_i-v_i};$$
$$0 \leqslant u \leqslant 1/\theta_i, \quad \theta_i > 0.$$

Case 1. $\theta_1 > \theta_2$.

$$\int_0^s f_2(t)\,dt = \int_0^s \frac{n_2!\,\theta_2}{(v_2-1)!\,(n_2-v_2)!}(\theta_2 t)^{v_2-1}\sum_{j=0}^{n_2-v_2}(-\theta_2 t)^j\binom{n_2-v_2}{j}dt$$

$$= \frac{n_2!}{(v_2-1)!}\sum_{j=0}^{n_2-v_2}\frac{(-1)^j(\theta_2 s)^{v_2+j}}{j!(n_2-v_2-j)!\,(v_2+j)}.$$

Since $f_2(s)$ is 0 for $s > 1/\theta_2 > 1/\theta_1$,

$$I = \int_0^{1/\theta_1}\left[\frac{n_1!}{(v_1-1)!\,(n_1-v_1)!}\,\theta_1(\theta_1 s)^{v_1-1}(1-\theta_1 s)^{n_1-v_1}\right]$$

$$\cdot\left[\frac{n_2!}{(v_2-1)!}\sum_{m=0}^{n_2-v_2}\frac{(-1)^m(\theta_2 s)^{v_2+m}}{m!(n_2-v_2-m)!\,(v_2+m)}\right]ds$$

$$I = \frac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!}\,\theta_1\sum_{m=0}^{n_2-v_2}\frac{(-1)^m\int_0^{1/\theta_1}(\theta_1 s)^{v_1-1}(1-\theta_1 s)^{n_1-v_1}(\theta_2 s)^{v_2+m}\,ds}{m!(n_2-v_2-m)!\,(v_2+m)(n_1-v_1)!}$$

Let

$$I_1 = \int_0^{1/\theta_1}(\theta_1 s)^{v_1-1}(1-\theta_1 s)^{n_1-v_1}(\theta_2 s)^{v_2+m}\,ds.$$

This is integrated by expanding the center term, resulting in

$$I_1 = \int_0^{1/\theta_1}\sum_{j=0}^{n_1-v_1}(-1)^j(\theta_1 s)^j\frac{(n_1-v_1)!}{j!(n_1-v_1-j)!}\,\theta_1^{v_1-1}\theta_2^{v_2+m}s^{v_2+v_2+m-1}\,ds \qquad (2)$$

$$= \sum_{j=0}^{n_1-v_1}\frac{(-1)^j(n_1-v_1)!}{j!(n_1-v_1-j)!}\,\theta_1^{-1}\left(\frac{\theta_2}{\theta_1}\right)^{v_2+m}\frac{1}{j+m+v_1+v_2},$$
$$j+m+v_1+v_2 \neq 0.$$

Upon substituting Eq. (2) into (1),

$$I = \frac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!}$$

$$\sum_{m=0}^{n_2-v_2}\left\{\frac{(-1)^m\left(\frac{\theta_2}{\theta_1}\right)^{v_2+m}}{m!(n_2-v_2-m)!\,(v_2+m)}\sum_{j=0}^{n_1-v_1}\frac{(-1)^j\frac{1}{j+m+v_1+v_2}}{j!(n_1-v_1-j)!}\right\}.$$

But note the identity[10]

$$\sum_{j=0}^{P} \frac{(-1)^j}{j!(P-j)!\,(a+j)} = \left[\prod_{j=0}^{P}(a+j)\right]^{-1} = \frac{(a-1)!}{(a+P)!}$$

$$I = \frac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!} \sum_{m=0}^{n_2-v_2} \frac{(-1)^m \left(\frac{\theta_2}{\theta_1}\right)^{v_2+m}}{m!(n_2-v_2-m)!\,(v_2+m)} \cdot \frac{(v_1+v_2+m-1)!}{(v_2+n_1+m)!}. \tag{3}$$

Equation (3) is the desired result.

Case 2. $\theta_1 < \theta_2$. The integral for Case 2 is identical to Case 1 except for interchange of class subscripts. It then results:

$$P_e = \begin{cases} \dfrac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!} \displaystyle\sum_{m=0}^{n_2-v_2} \dfrac{(-1)^m \left(\frac{\theta_2}{\theta_1}\right)^{v_2+m}}{m!(n_2-v_2-m)!} \dfrac{(v_1+v_2+m-1)!}{(v_2+m)(v_2+n_1+m)!}; & \theta_1 > \theta_2 \\ 1 - \dfrac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!} \displaystyle\sum_{m=0}^{n_1-v_1} \dfrac{(-1)^m \left(\frac{\theta_1}{\theta_2}\right)^{v_1+m}}{m!(n_1-v_1-m)!} \dfrac{(v_1+v_2+m-1)!}{(v_1+m)(v_1+n_2+m)!}; & \theta_1 < \theta_2. \end{cases}$$

APPENDIX III

We wish to evaluate I for the special case when $n_1 = n_2 = n$, $v_1 = v_2 = v$, $P_1W_1 > P_2W_2$ and $n \to \infty$, where

$$I = \begin{cases} \dfrac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!} \displaystyle\sum_{j=0}^{n_2-v_2} \dfrac{(-1)^j \left(\frac{\theta_2}{\theta_1}\right)^{v_2+j} (v_2+v_1+j-1)!}{j!(n_2-v_2-j)!\,(v_2+j)(v_2+n_1+j)!}; & \theta_1 > \theta_2 \quad \text{(1a)} \\ 1 - \dfrac{n_1!\,n_2!}{(v_1-1)!\,(v_2-1)!} \displaystyle\sum_{j=0}^{n_1-v_1} \dfrac{(-1)^j \left(\frac{\theta_1}{\theta_2}\right)^{v_1+j} (v_1+v_2+j-1)!}{j!(n_1-v_1-j)!\,(v_1+j)(v_1+n_2+j)!}; & \theta_1 < \theta_2 \quad \text{(1b)} \end{cases}$$

[10] I. J. Schwatt, "An Introduction to Operations With Series," 2nd ed. of 1924, 1st ed., p. 129. Chelsea, New York.

and

$$\theta_i = \frac{P_i v_i W_i}{n_i + 1}.$$

We will denote $I_\infty = \lim_{n\to\infty} I$. Since

$$\theta_1 = \frac{P_1 v_1 W_1}{n_1 + 1} > \frac{P_2 v_2 W_2}{n_2 + 1} = \theta_2,$$

we evaluate Eq. (1a).

$$I_\infty = \lim_{n\to\infty} \left\{ \left(\frac{n!}{(v-1)!}\right)^2 \sum_{j=0}^{n-v} \frac{(-1)^j \left(\frac{\theta_2}{\theta_1}\right)^{v+j} (2v-1+j)!}{j!(n-v-j)!\,(v+j)(v+n+j)!} \right\}$$

For n large but fixed, the general term of the sum is driven to 0 as j increases. Hence, let

$$\frac{n!}{(n+v+j)!} = n^{-v-j}[1 + o(n)]$$

$$\frac{n!}{(n-v-j)!} = n^{v+j}[1 + o(n)],$$

where $\lim_{n\to\infty} |\, o(n)| = 0$.

Then

$$I = \frac{1}{[(v-1)!]^2} \sum_{j=0}^{n-v} (-1)^j \left(\frac{\theta_2}{\theta_1}\right)^{v+j} \frac{(2v+j-1)}{(v+j)j!} [1 + O_1(n)][1 + O_2(n)]$$

and

$$I_\infty = \frac{1}{[(v-1)!]^2} \sum_{j=0}^{\infty} (-1)^j \left(\frac{\theta_2}{\theta_1}\right)^{v+j} \frac{(2v+j-1)}{(v+j)j!}.$$

We now obtain I_∞ in closed form. Since

$$\frac{r^{j+v}}{v+j} = \int_0^r \xi^{j+v-1} d\xi, \quad v > 0, \qquad \text{where} \quad r = \frac{\theta_2}{\theta_1} = \frac{P_2 W_2}{P_1 W_1}$$

$$I_\infty = \frac{1}{[(v-1)!]^2} \int_0^r \xi^{v-1} \sum_{j=0}^{\infty} \frac{(-1)^j (2v+j-1)!}{j!} \xi^j \, d\xi.$$

But since[11]

$$\frac{1}{(1+\xi)^{\mathscr{P}}} = \sum_{m=0}^{\infty} (-1)^m \frac{(\mathscr{P}+m-1)!}{m!(\mathscr{P}-1)!}\,\xi^m,$$

$$I_\infty = \frac{(2v-1)!}{[(v_1-1)!]^2}\int_0^r \frac{\xi^{v-1}}{(1+\xi)^{2v}}\,d\xi.$$

If $Z = 1 + \xi$, then

$$I_\infty = \frac{(2v-1)!}{[(v-1)!]^2}\int_1^{1+r} Z^{-2v}\sum_{j=0}^{v-1}(-1)^{v-1+j}\binom{v-1}{j}Z^j\,dZ$$

$$I_\infty = \frac{(2v-1)!}{[(v-1)!]^2}\sum_{j=0}^{v-1}(-1)^{v-1+j}\binom{v-1}{j}\left[\frac{1-(1+r)^{j-2v+1}}{-j+2v-1}\right].$$

Hence,

$$I_\infty = \frac{r}{1+r}, \qquad v = 1$$

$$= \frac{3r^2+r^3}{(1+r)^3}, \qquad v = 2$$

$$= \frac{10r^3+5r^4+r^5}{(1+r)^5}, \qquad v = 3$$

or

$$I_\infty = \left(\frac{1}{1+r}\right)^{2v-1}\sum_{j=v}^{2v-1}\binom{2v-1}{j}(r)^j, \qquad \text{where} \quad r = \frac{P_2W_2}{P_1W_1} = \frac{\theta_2}{\theta_1}.$$

RECEIVED: January 29, 1969; revised: September 19, 1969; revised: November 20, 1969

REFERENCES

1. D. O. LOFTSGAARDEN AND C. P. QUESENBURY, A nonparametric estimate of a multivariable density function, *Ann. Math. Statist.* **36** (1965), 1049–1151.
2. T. M. COVER AND P. E. HART, Nearest neighbor Pattern classification, *IEEE Trans.* **IT-13**, No. 1 (1967), 21–27.

[11] Mangulis, V., Handbook of Series for Scientists and Engineers, Academic Press, 1965, New York/London.

3. E. FIX AND J. L. HODGES, JR., "Discriminatory Analysis; Small Sample Performance," USAF School of Aviation Medicine, Randolph Field, Texas, Project 21–49–004, Report No. 11, under Contract No. AF41(148)-31, August, 1952.
4. S. S. WILKS, Determination of sample sizes for setting tolerance limits, *Ann. Math. Statist.* **12** (1941), 91–96.
5. J. H. B. KEMPERMAN, Generalized tolerance limits, *Ann. Math. Statist.* **27** (1956), 180–186.
6. D. A. S. FRASER AND I. GUTTMAN, Tolerance regions, *Ann. Math. Statist.* **27** (1956), 162–179.
7. D. A. S. FRASER, "Nonparametric Methods in Statistics," Chap. 4.3, Wiley, New York, 1957.
8. S. S. WILKS, "Mathematical Statistics," Chap. 18, Wiley, New York, 1962.
9. J. W. TUKEY, Nonparametric estimation, II. Statistically equivalent blocks and tolerance regions—the continuous case, *Ann. Math. Statist.* **18** (1947), 529–539.
10. E. A. PATRICK, "Distribution Free, Minimum Conditional Risk Learning Systems" Purdue Technical Report, TR–EE66–18, November 1966.
11. E. A. PATRICK, K. FUKUNAGA, D. R. ANDERSON, F. P. FISCHER, II., L. Y. L. SHEN, Final Report, Part II., Experimental Results, "Supervised and Unsupervised Adaptive System for Submarine Detection (*U*)," Purdue University TR–EE68–22, June, 1968.
12. E. A. PATRICK AND J. P. COSTELLO, "Bayes Related Solutions to Unsupervised Estimation," Proceedings of the 1969 National Electronics Conference, December 1969.