capability upon stimulus predictability," *Vis. Res.*, vol. 6, pp. 707–716, 1966.

[20] P. Morasso, E. Bizzi, and J. Dichgans, "Adjustment of saccade characteristics during head movements," *Exp. Brain Res.*, vol. 16, pp. 492–500, 1973.

[21] P. Morasso, G. Sandini, and V. Tagliasco, "Plasticity in the eye head coordination system," in *Aspects of Neural Plasticity*, Vital-Durand and Jeannerod, Eds., INSERM, vol. 43, pp. 83–94, 1975.

[22] P. Morasso, "EGOS: A minioperating system for management of experiments in neurophysiology and behavioural sciences," *Proc. IFAC-IFIP Workshop on Real-Time Programming*, Budapest, pp. 259–263, 1974.

[23] C. Prablanc and M. Jeannerod, "Corrective saccades: Dependence on retinal reafferent signals," *Vision Res.*, vol. 15, pp. 465–469, 1975.

[24] D. A. Robinson, "The mechanics of human saccadic eye movement," *J. Physiol.*, vol. 174, pp. 245–264, 1964.

[25] ——, "The oculomotor control system: A review," *Proc. IEEE*, vol. 56, pp. 1032–1048, 1968.

[26] ——, "Models of saccadic eye movement control system," *Kibernetik*, vol. 14, pp. 71–83, 1973.

[27] L. Stark, G. Vossius, and L. R. Young, "Predictive control of eye tracking movements," *IEEE Trans. Human Factors in Electronics*, vol. HFE-3, pp. 52–57, 1962.

[28] N. Sugie and M. Wakakuwa, "Visual target tracking with active head rotation," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-6, pp. 103–108, 1970.

[29] N. Sugie and G. Melvill-Jones, "A model of eye movements induced by head rotation," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-1, pp. 251–260, 1971.

[30] N. Sugie, "A model of predictive control in visual target tracking," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-1, pp. 2–7, 1971.

[31] L. R. Young, "A sampled data model for eye tracking movements," Sc.D. thesis, Dept. Aeronautics and Astronautics, MIT, Cambridge, 1962.

[32] L. R. Young and L. Stark, "Variable feedback experiments testing a sampled data model for eye tracking movements," *IEEE Trans. Human Factors in Electronics*, vol. HFE-4, pp. 52–57, 1962.

[33] R. Zaccaria, "BSSP: Biological systems simulation program," Tech. Rep., Istituto di Elettrotecnica Genova, vol. TR 1/75, 1975.

Nonparametric Bayes Risk Estimation for Pattern Classification

ZEN CHEN AND KING-SUN FU, FELLOW, IEEE

Abstract—**The performance of a pattern classification system is often evaluated based on the risk committed by the classification procedure. The minimum attainable risk is the Bayes risk. Therefore, the Bayes risk can be used as a measure of the intrinsic complexity of the system, and it also serves as a reference of the optimality measure of a classification procedure. There are many practical situations in which the nonparametric methods may have to be called upon to estimate the Bayes risk. One of the nonparametric methods is via the probability density estimation technique. The convergence properties of this estimation technique are studied under fairly general assumptions. In the computer experiments reported, the estimate of the Bayes risk is taken as the sample mean of the density estimate by making use of the leave-one-out method. The probability density estimate used is the one proposed by Loftsgaarden and Quesenberry. This estimate is shown to be, in general, superior to the risk associated with a Bayes-like decision rule based on the error-counting scheme. This estimate is also compared experimentally with the risk estimate associated with the nearest neighbor rule.**

Manuscript received December 13, 1976; revised March 21, 1977. This work was supported by the Air Force Office of Scientific Research under Grant 74-2661.

Z. Chen is with the Institute of Computer Science, National Chiao Tung University, Hsin-Chu, Taiwan.

K. S. Fu is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

I. Introduction

ASSUME there exists a class of conditional probability densities $F = \{f_1, f_2, \cdots, f_M\}$ in a probability space $\{S,B,P\}$, where S is the sample space, B is a σ-algebra of subsets of S, and P is a probability measure on B. Let $\eta_1, \eta_2, \cdots, \eta_M$, $\eta_i \geq 0$, $\sum_{i=1}^{M} \eta_i = 1$, be the prior probabilities of occurrence of the M pattern classes. Also let $L(i,j)$ be the loss incurred by classifying a sample from class i into class j. A pattern classification procedure is to assign a new sample in the sample space (usually in the form of measurement vectors) to one of the M pattern classes. The performance of a pattern classification system is properly evaluated based on the risk (i.e., the expected value of the loss due to misclassification) committed by the classification procedure. The minimum attainable risk is the Bayes risk. Therefore, the Bayes risk can be used as a measure of the intrinsic complexity of the system, and it also serves as a reference of an optimality measure of a procedure.

The Bayes risk is a function of *a priori* probabilities and the underlying conditional probability densities. In the case where $\{\eta_i\}$ and $\{f_i\}$ are known completely, it is well known

[21] that the following randomized decision rule d^B attains the Bayes risk:

$$d_j^B(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^{M} \eta_i L(i,j) f_i(x) \\ & \qquad < \min_{k \neq j} \sum_{i=1}^{M} \eta_i L(i,k) f_i(x) \\ 0, & \text{if } \sum_{i=1}^{M} \eta_i L(i,j) f_i(x) \\ & \qquad > \min_{k \neq j} \sum_{i=1}^{M} \eta_i L(i,k) f_i(x) \\ \alpha_j, & \text{if } \sum_{i=1}^{M} \eta_i L(i,j) f_i(x) \\ & \qquad = \min_{k \neq j} \sum_{i=1}^{M} \eta_i L(i,k) f_i(x) \end{cases} \tag{1}$$

where x is the observable measurement of the new sample X, and

$$\sum_{j=1}^{M} \alpha_j = 1, \qquad \alpha_j \geq 0.$$

The Bayes risk R_B is given by

$$R_B = \sum_{i=1}^{M} \eta_i E_i \gamma_B(i,X),$$

$$\gamma_B(i,X) = \sum_{j=1}^{M} L(i,j) d_j^B(X),$$

where $\gamma_B(i,x)$ is the risk associated with class i committed by the decision rule d^B, given that the random vector X takes on the value x, and E_i is the expectation taken over S with respect to $f_i \in F$.

In the real world, there is often a lack of the exact knowledge of $\{\eta_i\}$ and $\{f_i\}$; instead only partial information is available. For instance, there are situations in which only the parametric forms of the underlying distributions and/or a set of correctly classified samples from the distributions are known. Based on this partial information, parametric and nonparametric methods have been studied by many researchers to solve the pattern classification problem. Nonparametric methods are used under the condition that no parametric forms of underlying distributions are known or can be assumed, [1]–[11].

In this paper the focus is placed on the nonparametric Bayes risk estimation via the sample mean of an estimator of the conditional Bayes risk which, in turn, employs the density estimation technique. Various asymptotic properties of the above conditional risk estimator are studied under fairly general assumptions. The nonparametric Bayes risk estimation is implemented with the given correctly classified samples on a digital computer. The experimental results are then discussed.

II. Nonparametric Classification Procedure with Density Estimation

In classifying a new sample into one of the M possible pattern classes, there are two categories of nonparametric classification procedures. On the one hand, there are procedures which do not involve the use of any form of the underlying probability densities. Under this category there are i) the nearest neighbor decision rule [1], [2], [4]–[7]; ii) classification procedures based on statistically equivalent blocks [3], [18]; iii) the classification by linear or piecewise linear discriminant functions [19]; and others. On the other hand, there are procedures which employ density estimation techniques [8]–[10]. These procedures are conceptually simple and are analogous to those parametric methods in statistical decision theory. To facilitate later discussion, one of the general forms of these procedures is given here.

Assume that the density functions $f_i \in F$, $i = 1, 2, \cdots, M$, are estimated from the training sample sets by making use of some density estimation technique. Let $\hat{f}_{i,n_i}(x)$ denote the estimate of $f_i(x)$, $i = 1, 2, \cdots, M$ from a set of training samples $\{X_i^{(i)}, \cdots, X_{n_i}^{(i)}\}$, $i = 1, 2, \cdots, M$. Let $\boldsymbol{X}_n \triangleq \{X_1^{(1)}, \cdots, X_{n_1}^{(1)}, \cdots, X_1^{(M)}, \cdots, X_{n_M}^{(M)}\} \triangleq \{X_1, X_2, \cdots, X_n\}$, $n_1 + \cdots + n_M = n$, and let $\hat{\eta}_i = n_i/n$, $i = 1, 2, \cdots, M$. Based on these estimates, a decision rule, denoted by $d^0(x)$, which is directed by the Bayes rule, is defined as follows:

$$d_j^0(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^{M} \hat{\eta}_i L(i,j) \hat{f}_{i,n_i}(x) \\ & \qquad < \min_{k \neq j} \sum_{i=1}^{M} \hat{\eta}_i L(i,k) \hat{f}_{i,n_i}(x) \\ 0, & \text{if } \sum_{i=1}^{M} \hat{\eta}_i L(i,j) \hat{f}_{i,n_i}(x) \\ & \qquad > \min_{k \neq j} \sum_{i=1}^{M} \hat{\eta}_i L(i,k) \hat{f}_{i,n_i}(x) \\ \alpha_i, & \text{if } \sum_{i=1}^{M} \hat{\eta}_i L(i,j) \hat{f}_{i,n_i}(x) \\ & \qquad = \min_{k \neq j} \sum_{i=1}^{M} \hat{\eta}_i L(i,k) \hat{f}_{i,n_i}(x) \end{cases} \tag{2}$$

where $\alpha_j \geq 0$, $j = 1, 2, \cdots, M$, and $\sum_{j=1}^{M} \alpha_j = 1$.

III. Estimation of Bayes Risk

Once the classification procedure is devised, the performance is mainly evaluated by the misclassification committed by the procedure. In certain cases the estimation of misclassification can be related to the Bayes risk and is therefore used to estimate the latter. It was shown [4] that in a two-class problem, the risk of the 1-nearest neighbor rule R with the (0,1)-loss function is related to the Bayes risk R^* by $R^* \leq R \leq 2R^*(1 - R^*)$. Let S_n/n be an estimate of R, then the interval $[(1 - \sqrt{1 - 2S_n/n})/2, S_n/n]$ is an estimate of R^*.

The risk associated with the decision rule d^0 given by (2) using the error-counting scheme was indicated to converge to the Bayes risk in quadratic mean [9].

A different estimation of the Bayes risk can be built upon an estimator of the conditional Bayes risk. The conditional Bayes risk $\gamma_B(x)$ corresponding to (1) is given by

$$\gamma_B(x) = \min_{j \in \{1, \cdots, M\}} \left\{ \sum_{i=1}^{M} L(i,j)\rho_i(x) \right\}$$

$$\rho_j(x) = \{\eta_j f_j(x)\} \Big/ \left\{ \sum_{i=1}^{M} \eta_i f_i(x) \right\}.$$

Now define

$$\hat{\rho}_{j,n}(x \mid X_n) = \{\hat{\eta}_j \hat{f}_{j,n_j}\} \Big/ \left\{ \sum_{i=1}^{M} \hat{\eta}_i \hat{f}_{i,n_i}(x) \right\}$$

and

$$\gamma_n^0(x \mid X_n) = \min_{j \in \{1, \cdots, M\}} \left\{ \sum_{i=1}^{M} L(i,j)\hat{\rho}_{i,n}(x \mid X_n) \right\}.$$

Notice that $\hat{\rho}_{j,n}(x \mid X_n)$ and $\gamma_n^0(x \mid X_n)$ are conditioned on X_n and, therefore, are random variables.

It will be shown that $\gamma_n^0(x \mid X_n)$ is a consistent estimator of $\gamma_B(x)$. Consequently, R_B can be inferred by an estimator of $E_X \gamma_n^0(X \mid X_n)$.

IV. Asymptotic Properties of $\gamma_n^0(X \mid X_n)$

Before the asymptotic properties of $\gamma_n^0(X \mid X_n)$ are studied, some assumptions and notations will be introduced first. In the following it will be assumed that the conditional probability densities f_i, $i = 1, 2, \cdots, M$, are absolutely continuous and bounded from above, and that their estimates $\hat{f}_{i,n_i}(x)$, $i = 1, 2, \cdots, M$, are nonnegative. In addition, assume $\hat{f}_{i,n_i}(x) \xrightarrow{P} f_i(x)$, $i = 1, 2, \cdots, M$; namely, $\hat{f}_{i,n_i}(x)$ converges to $f_i(x)$ in probability for $i = 1, 2, \cdots, M$. Also assume that the loss functions $L(i,j)$, $i, j = 1, 2, \cdots, M$, are nonnegative and finite. Let the notation $n_i \to \infty$ indicate $n_i \to \infty$, for $i = 1, 2, \cdots, M$. Besides, $E_{X_n} \gamma_n^0(x \mid X_n)$ means the expectation is taken with respect to all $X_1, X_2, \cdots, X_n$. Analogous interpretations apply to $E_X \gamma_n^0(X \mid X_n)$ and $E_{X_n} E_X \gamma_n^0(X \mid X_n)$. Finally define

$$R_n^0 = E_{X_n}\{E_X \gamma_n^0(X \mid X_n)\} = E_{X_n}\left\{ \int \gamma_n^0(x \mid X_n) \left[\sum_{i=1}^{M} \eta_i f_i(x) \right] dx \right\}.$$

Lemma 1:

$$\text{i)} \quad \gamma_n^0(x \mid X_n) \xrightarrow{P} \gamma_B(x)$$

and

$$\text{ii)} \quad \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n) = \gamma_B(x).$$

Proof: Since $\hat{\eta}_i \to \eta_i$, $i = 1, 2, \cdots, M$, almost everywhere (a.e.) by the law of large numbers [17], this, together with $\hat{f}_{i,n_i}(x) \xrightarrow{P} f_i(x)$, $i = 1, 2, \cdots, M$, implies $\hat{\rho}_{j,n}(x \mid X_n) \xrightarrow{P} \rho_j(x)$, $j = 1, 2, \cdots, M$, and, therefore, $\gamma_n^0(x \mid X_n) \xrightarrow{P} \gamma_B(x)$.

Furthermore, for every n ($\forall n$), with probability 1,

$$|\gamma_n^0(x \mid X_n)| = \left| \min_{j \in \{1, \cdots, M\}} \left\{ \sum_{i=1}^{M} L(i,j)\hat{\rho}_{i,n}(x \mid X_n) \right\} \right|$$
$$\leq \left| \min_{j} \left\{ \sum_{i=1}^{M} \bar{L}\hat{\rho}_{i,n}(x \mid X_n) \right\} \right| = \bar{L} < \infty$$

where $\bar{L}$ is the maximum value of $L(i,j)$, $i, j = 1, \cdots, M$.

By Lebesgue's dominated convergence theorem,

$$\lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n) = \gamma_B(x)$$

$\forall x$ except for a set of points with zero probability measure.

Theorem 1: R_n^0 converges to the Bayes risk R_B in the ordinary sense, as $n_i \to \infty$.

Proof: Since $E_{X_n} \gamma_n^0(x \mid X_n) < \bar{L}$ almost everywhere and $\forall n$ by Lebesque's dominated convergence theorem,

$$\begin{aligned} \lim_{n_i \to \infty} R_n^0 &= \lim_{n_i \to \infty} E_X E_{X_n} \gamma_n^0(X \mid X_n) \\ &= E_X \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(X \mid X_n) \\ &= E_X \gamma_B(x) = R_B. \end{aligned}$$

Theorem 2:

$$\gamma_n^0(x \mid X_n) \xrightarrow{P} E_{X_n} \gamma_n^0(x \mid X_n).$$

Proof:

$$\begin{aligned} &|\gamma_n^0(x \mid X_n) - E_{X_n} \gamma_n^0(x \mid X_n)| \\ &\quad = |\gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n) - E_{X_n} \gamma_n^0(x \mid X_n) \\ &\qquad + \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \\ &\quad \leq |\gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \\ &\qquad + |E_{X_n} \gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)|. \end{aligned}$$

Given $\varepsilon > 0$, we can find $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\varepsilon = \varepsilon_1 + \varepsilon_2$. By Lemma 1,

$$\lim_{n_i \to \infty} \Pr\{|\gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \leq \varepsilon_2\} = 1$$

and there exists m_i, $i = 1, 2, \cdots, M$, if $n_i \geq m_i$, $i = 1, 2, \cdots, M$,

$$|E_{X_n} \gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \leq \varepsilon_1$$

$$\begin{aligned} &\lim_{n_i \to \infty} \Pr\{|\gamma_n^0(x \mid X_n) - E_{X_n} \gamma_n^0(x \mid X_n)| \leq \varepsilon\} \\ &\quad \geq \lim_{n_i \to \infty} \Pr\{|\gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \\ &\qquad + |E_{X_n} \gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \leq \varepsilon\} \\ &\quad \geq \lim_{n_i \to \infty} \Pr\{|\gamma_n^0(x \mid X_n) - \lim_{n_i \to \infty} E_{X_n} \gamma_n^0(x \mid X_n)| \leq \varepsilon - \varepsilon_1\} \\ &\quad = 1 \end{aligned}$$

i.e., $\gamma_n^0(x \mid X_n) \xrightarrow{P} E_{X_n} \gamma_n^0(x \mid X_n)$.

Corollary 1: $\gamma_n^0(x \mid X_n) \to E_{X_n} \gamma_n^0(x \mid X_n)$ in the kth mean ($k > 0$).

Proof: Since $|\gamma_n^0(x \mid X_n)| \leq \bar{L} < \infty$, a.e., $\forall n$, the convergence in probability implies the convergence in the kth mean [17]. Therefore, Corollary 1 follows from Theorem 2.

Theorem 3: $E_X\gamma_n^0(X\,|\,X_n) \xrightarrow{P} E_X\{E_{X_n}\gamma_n^0(X\,|\,X_n)\}$.

Proof: For any $\varepsilon > 0$, by the Markov inequality [17],

$$\lim_{n_i\to\infty} \Pr\{|E_X\gamma_n^0(X\,|\,X_n) - E_X(E_{X_n}\gamma_n^0(X\,|\,X_n))| \ge \varepsilon\}$$

$$\le \lim_{n_i\to\infty} \{E_{X_n}|E_X\gamma_n^0(X\,|\,X_n) - E_X(E_{X_n}\gamma_n^0(X\,|\,X_n))|\}/\varepsilon$$

$$\le \lim_{n_i\to\infty} \left\{E_{X_n}\left(c\int |\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)|\,dx\right)\right\}\Big/\varepsilon$$

where c is a constant such that

$$f(x) = \sum_{i=1}^{M} \eta_i f_i(x) \le c$$

for every x except for a set of points with zero probability measure, since

$$|E_X\gamma_n^0(X\,|\,X_n) - E_X E_{X_n}\gamma_n^0(X\,|\,X_n)|$$

$$\le E_X|\gamma_n^0(X\,|\,X_n) - E_{X_n}\gamma_n^0(X\,|\,X_n)|$$

$$= \int |\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)|\,f(x)\,dx$$

$$\le \int c\,|\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)|\,dx.$$

Now $E_{X_n}|\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)| \le c' < \infty$, a.e., $\forall n$, by Corollary 1,

$$\lim_{n_i\to\infty} E_{X_n}|\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)| = 0.$$

By Lebesgues's dominated convergence theorem,

$$\lim_{n_i\to\infty} \Pr\{|E_X\gamma_n^0(X\,|\,X_n) - E_X(E_{X_n}\gamma_n^0(X\,|\,X_n))| \ge \varepsilon\}$$

$$\le c\left\{\int \lim_{n_i\to\infty} E_{X_n}|\gamma_n^0(x\,|\,X_n) - E_{X_n}\gamma_n^0(x\,|\,X_n)|\,dx\right\}\Big/\varepsilon = 0$$

i.e.,

$$E_X\gamma_n^0(X\,|\,X_n) \xrightarrow{P} E_X\{E_{X_n}\gamma_n^0(X\,|\,X_n)\}.$$

Corollary 2: $E_X\gamma_n^0(X\,|\,X_n) \to E_X\{E_{X_n}\gamma_n^0(X\,|\,X_n)\}$ in the kth mean.

Theorem 4: $E_X\gamma_n^0(X\,|\,X_n) \xrightarrow{P} R_B$.

Proof: From above,

$$R_B = \lim_{n_i\to\infty} E_{X,X_n}\gamma_n^0(X\,|\,X_n)$$

and

$$E_X\gamma_n^0(X\,|\,X_n) \xrightarrow{P} E_X\{E_{X_n}\gamma_n^0(X\,|\,X_n)\}.$$

By a technique similar to the one used in the proof of Theorem 2, it can be shown that

$$E_X\gamma_n^0(X\,|\,X_n) \xrightarrow{P} R_B.$$

Since $|E_X\gamma_n^0(X\,|\,X_n)| \le \bar{L} < \infty$, a.e., $\forall n$, the following corollary can be shown.

Corollary 3: $E_X\gamma_n^0(X\,|\,X_n) \to R_B$ in the kth mean. In particular,

$$0 \le E_{X_n}(E_X\gamma_n^0(X\,|\,X_n)) - R_B \le E_{X_n}|E_X\gamma_n^0(X\,|\,X_n) - R_B| \to 0$$

as $n_i \to \infty$ and $E_{X_n}|E_X\gamma_n^0(X\,|\,X_n) - R_B|^2 \to 0$ as $n_i \to \infty$ $\forall i$.

The previous theorems and corollaries are the foundation for $\gamma_n^0(X\,|\,X_n)$ to be used in the Bayes risk estimation. The expectation of $\gamma_n^0(X\,|\,X_n)$ is shown to converge to the Bayes risk in probability as well as in the kth mean, in contrast to the convergence of the risk of the nearest neighbor rule which is only to a bound on the Bayes risk. This result indicates the use of a sample mean of $\gamma_n^0(X\,|\,X_n)$ as a desirable estimation of the Bayes risk. The empirical comparison of this estimator with the other estimators will be given in the next section.

V. Utilization of Given Samples in Estimation

In order to estimate the risk by using the correctly classified samples, two things must be decided. One is to choose a probability density estimation technique and the other is to decide on a method to effectively utilize the available labeled samples to carry out the estimation scheme. As far as the first problem is concerned, the density estimator proposed originally by Parzen [13] and extended later by Cacoullos [14], the one by Murthy [15], and the one by Loftsgaarden and Quesenberry [12] all meet the consistency requirement of the density estimation. It is the latter one which was employed in the computer experiments reported.

In utilizing given labeled samples to carry out the risk estimation, there are mainly three methods. They are i) the resubstitution method; ii) the holdout method, or H method; and iii) the leave-one-out method, or the U method [16]. Generally speaking, the first method gives an overly optimistic estimate, while the second method gives an overly pessimistic estimate. The third method yields an estimate with a small amount of bias compared with those of the previous two methods, although it suffers from requiring more computation time. It is the third method which was used in the computer experiments.

The application of the leave-one-out method to the estimation of the Bayes risk discussed previously leads to an estimator $\hat{R}_n^0(X_n)$ which is given by

$$\hat{R}_n^0(X_n) = 1/n \sum_{i=1}^{M} \gamma_{n-1}^0(X_i|X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n).$$

The use of $\hat{R}_n^0(X_n)$ in estimating the Bayes risk is justified by the following convergence theorem.

Theorem 5: $\hat{R}_n^0(X_n)$ converges to the Bayes risk R_B in probability as well as in the kth mean ($k > 0$).

From Lemma 1, it can be shown that

$$\gamma_{n-1}^0(X_i|X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$$

$$\triangleq \gamma_{n-1}^0(X_i\,|\,X_{n-1}) \xrightarrow{P} \gamma_B(X_i).$$

By the Markov inequality,

$$\Pr\left\{\left|\sum_{i=1}^{n}\gamma_{n-1}^{0}(X_i\,|\,X_{n-1})-\sum_{i=1}^{n}\gamma_B(X_i)\right|\geq\varepsilon\right\}$$

$$\leq\left\{E\left|\sum_{i=1}^{n}\gamma_{n-1}^{0}(X_i\,|\,X_{n-1})-\sum_{i=1}^{n}\gamma_B(X_i)\right|\right\}\Big/\varepsilon$$

$$\leq(1/\varepsilon)\left\{\sum_{i=1}^{n}E\left|\gamma_{n-1}^{0}(X_i\,|\,X_{n-1})-\gamma_B(X_i)\right|\right\}.$$

By the convergence theorem in the kth mean [17],

$$(1/n)\sum_{i=1}^{n}\gamma_{n-1}^{0}(X_i\,|\,X_{n-1})\xrightarrow{P}1/n\sum_{i=1}^{n}\gamma_B(X_i).$$

Because the X_i are i.i.d., the $\gamma_B(X_i)$ are also i.i.d. By Bernoulli's law,

$$(1/n)\sum_{i=1}^{n}\gamma_B(X_i)\xrightarrow{P}E_X\gamma_B(X)=R_B.$$

Thus

$$(1/n)\sum_{i=1}^{n}\gamma_{n-1}^{0}(X_i\,|\,X_{n-1})\rightarrow R_B$$

in probability as well as in the kth mean.

An interesting remark is in order. Experimental results indicate that $\hat{R}_n^0(X_n)$ has a smaller variance than does the Bayes risk estimate by the risk associated with the classification procedure obtained from the error counting method. The reason may lie in the fact that $\hat{R}_n^0(X_n)$ is a smoother function compared with the error counting risk estimate. Therefore, $\hat{R}_n^0(X_n)$ converges to the Bayes risk more rapidly.

VI. Computer Experiments

The estimation of the Bayes risk discussed above is implemented on a digital computer. In the following, the constant $\{k_n\}$ is referred to as the sequence of positive integers of the Loftsgaarden and Quesenberry estimator of the density estimation. Let $\hat{R}_n^0$, $\hat{R}_E$, and $\hat{R}_k$ be the estimators of the Bayes risk by three different models, namely, those based on $E_X\gamma_n^0(X\,|\,X_n)$, the risk associated with the decision rule $d^0(x)$, and the k-NN decision rule, respectively. The data used in the experiments are bivariate Gaussian data $N(\mu_i,\Sigma_i)$, $i=1,2$, where

$$\mu_1=\begin{pmatrix}3.0\\-1.0\end{pmatrix}\qquad\mu_2=\begin{pmatrix}-3.0\\-1.0\end{pmatrix}$$

and

$$\Sigma_1=\Sigma_2=\begin{pmatrix}2.0&0\\0&2.0\end{pmatrix}.$$

Experiment 1: Five sets of samples were generated with $n_1=n_2=n=100$, 150, 200. $\hat{R}_n^0$ is obtained by the leave-one-out method and the holdout method. The holdout sample sizes corresponding to $n=100$, 150, and 200 are 25, 50, and 75, respectively. The results are shown in Table I. As the results indicate, the U method is better than the H method.

TABLE I
Average and Standard Deviation for $\hat{R}_n^0$ with $k_n=(n)^{0.55}$

n	$\hat{R}_n^0$			
	U Method		H Method	
	Avg.	SD	Avg.	SD
100	0.1363	0.0259	0.1472	0.0461
150	0.1275	0.0067	0.1406	0.0357
200	0.1248	0.0067	0.1357	0.0271

TABLE II
Average and Standard Deviation for $\hat{R}_n^0$ and $\hat{R}_E$ with $n_1=n_2=150$, $k_n=(n)^{\alpha_i}$

α_i	$\hat{R}_n^0$(U Method)		$\hat{R}_E$	
	Avg.	SD	Avg.	SD
0.45	0.1168	0.0111	0.0760	0.0134
0.50	0.1242	0.0113	0.0733	0.0139
0.55	0.1275	0.0067	0.0700	0.0139

TABLE III
Average and Standard Deviation for $\hat{R}_n^0$, $\hat{R}_E$, and $\hat{R}_3$ with $k_n=(n)^{0.55}$

$n_1=n_2$	$\hat{R}_n^0$(U Method)		$\hat{R}_E$		$\hat{R}_3$	
	Avg.	SD	Avg.	SD	Avg.	SD
50	0.1562	0.0126	0.060	0.0316	——	——
75	0.1320	0.0152	0.0653	0.0152	0.0627	0.2433
100	0.1363	0.0259	0.064	0.0297	0.079	0.0222
150	0.1275	0.0067	0.070	0.0139	0.0733	0.0238
200	0.1248	0.0067	0.066	0.0133	0.077	0.0124

Experiment 2: Five sets of samples were generated with $n_1=n_2=n=150$. $\hat{R}_n^0$ is obtained for three different values of k_n, i.e., $k_n=n^{\alpha_i}$, $i=1,2,3$, with $\alpha_1=0.45$, $\alpha_2=0.5$, and $\alpha_3=0.55$. The results are summarized in Table II. $\hat{R}_n^0$ is much closer to the true risk ($R_B=0.134$) than $\hat{R}_E$. For the sample sizes under consideration, $\hat{R}_n^0$ is superior to $\hat{R}_E$.

Experiment 3: For $n_1=n_2=25$, 50, 75, 100, 150, 200, five sets of training samples were generated. The three Bayes risk estimates $\hat{R}_n^0$, $\hat{R}_E$, and $\hat{R}_3$ are computed. The experimental results are shown in Table III and Fig. 1. We can find that $\hat{R}_n^0$ is better than $\hat{R}_E$ and $\hat{R}_3$. Both $\hat{R}_E$ and $\hat{R}_3$ are optimistic risk estimates.

It is important to know that the ability of $\hat{R}_n^0(X_n)$ in estimating the Bayes risk well relys on the appropriate choice of the constant k_n. An improper choice may lead to poor results [20].

VII. Conclusions

In this paper focus is placed on the nonparametric Bayes risk estimation. The estimate based on the conditional Bayes risk estimator is employed by making use of the density estimation. Various asymptotic properties of this estimate are studied under mild assumptions.

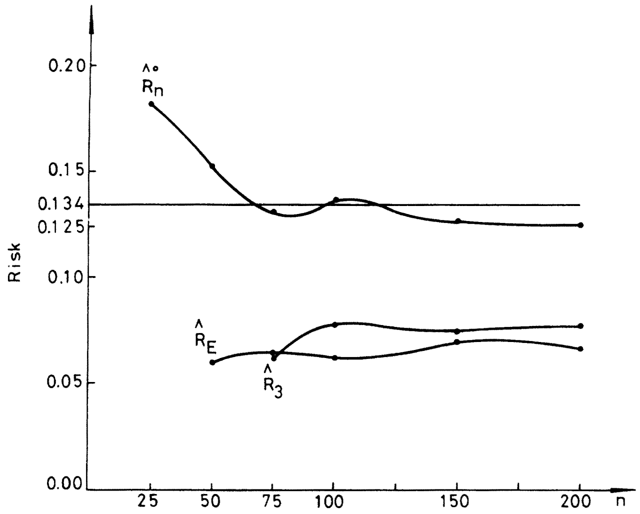

Fig. 1. Plot of Bayes risk estimates versus n (number of samples per class).

The computer implementation of the estimate of Bayes risk with the given samples by the leave-one-out method is given. The estimate converges to the Bayes risk in probability and in the kth mean. It was shown that the risk estimate based on the conditional Bayes risk is generally superior to the risk estimate associated with a Bayes-like decision rule d^0. For the Gaussian data used, $\hat{R}_n^0$ was shown to be closer to the Bayes risk than $\hat{R}_E$ and $\hat{R}_k$ were. However, a bad choice of constants in the estimate may cause the results to deteriorate.

The Bayes risk estimation discussed above requires no *a priori* information of the pattern underlying distribution. This point lends the method to a number of applications.

1) The Feature Selection Problem: Quite often in a pattern classification problem there are no clear rules for selecting an effective set of features. A good practice is to select a set of feature which yields the smallest value of the estimated Bayes risk. The set of features thus obtained generally will lead to an efficient classifier design.

2) The Measure of Separability of Clusters: In cluster analysis the measure of separability of the resultant clusters by means of the Bayes risk estimation can provide good insight into the structure of clusters.

References

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination," USAF School of Aviation Medicine, Randolph Field, TX, Project 21-49-004, Rep. 4, Contract AF41 (128)-31, Feb. 1951.

[2] ——, "Discriminatory analysis: Small sample performance," USAF School of Aviation Medicine, Randolph Field, TX, Project 21-49-004, Rep. 11, Aug. 1952.

[3] T. W. Anderson, "Some nonparametric multivariate procedures based on statistically equivalent blocks," in *Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic Press, 1966, pp. 5–27.

[4] T. M. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.

[5] D. W. Paterson, "Some convergence properties of a nearest neighbor decision rule," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 26–31, Jan. 1970.

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

[7] D. L. Wilson, "Asymptotic properties of nearest neighbor rule using edited data," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-2, pp. 408–421, July 1972.

[8] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya*, Ser. A, pt. 2-3, vol. 28, pp. 261–270, Sept. 1966.

[9] S. C. Fralick and R. W. Scott, "Nonparametric Bayes risk estimation," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 440–444, July 1971.

[10] E. A. Patric and F. P. Fischer, II, "Introduction to the performance of distribution free, minimum conditional risk learning systems," School of Electrical Engineering, Purdue Univ., Lafayette, IN, Tech. Rep. EE67-12, July 1967.

[11] T. Cover, "Learning in pattern recognition," in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic Press, 1969, pp. 111–132.

[12] D. O. Loftsgaarden and C. D. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, pp. 1049–1051, 1965.

[13] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[14] T. Cacoullos, "Estimation of a multivariate density," *Ann. Math. Statist.*, vol. 18, pp. 179–189, 1966.

[15] V. Murthy, "Estimation of probability density," *Ann. Math. Statist.*, vol. 36, pp. 1027–1031, 1965.

[16] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 715–725, Feb. 1968.

[17] M. Loeve, *Probability Theory*, third ed. Princeton, NJ: Van Nostrand, 1963.

[18] G. W. Beakley and F. B. Tuteur, "Distribution-free pattern verification using statistically equivalent blocks," *IEEE Trans. Comput.*, vol. C-21, pp. 1337–1347, Dec. 1973.

[19] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1972.

[20] Z. Chen, "Nonparametric methods for nonsupervised and supervised pattern recognition," Ph.D. dissertation, Purdue Univ., West Lafayette, IN, 1973.

[21] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press, 1967.