Bibliography on Estimation of Misclassification

GODFRIED T. TOUSSAINT

Abstract—**Articles, books, and technical reports on the theoretical and experimental estimation of probability of misclassification are listed for the case of correctly labeled or preclassified training data. By way of introduction, the problem of estimating the probability of misclassification is discussed in order to characterize the contributions of the literature.**

INTRODUCTION

ONE OF THE most important problems in pattern recognition is estimating the probability of misclassification. Before embarking on a description of the literature in this field it is proper to make a distinction between some of the various measures of probabilities of misclassification (error) usually considered.

1) The *optimal or Bayes probability of error*, denoted by P_e^B, is given by $P_e^B = 1 - \int \max_i \{P(X/C_i)P(C_i)\}\, dX$, where $P(X/C_i)$ and $P(C_i)$ are the class-conditional probability density function and the *a priori* probability of the ith class, respectively. This error probability results when one has complete knowledge of the probability density functions with which to construct the optimal decision rule and uses the Bayes decision rule. Knowledge of the underlying distributions could have resulted from observing an infinite number of independent labeled pattern samples. Thus the Bayes error rate can be thought of as the infinite-sample error.

2) The error probability that results when one has complete knowledge of the probability density functions and uses a decision rule other than the Bayes' rule is denoted by P_e^c.

It is clear that $P_e^c \geq P_e^B$, with equality holding when the given classifier is (Bayes) optimal. The term P_e will be used when no distinction is made between P_e^c and P_e^B, and it will be referred to as the "actual" error probability.

In practice one usually obtains a data set which is not only finite, but in fact quite small. Frequently no knowledge is available concerning the underlying distributions. In such situations one would, ideally, like to know what the resulting probability of error is going to be on future pattern samples when the classifier is trained, i.e., its parameters estimated, on the given data set.

3) Denote by $\hat{P}_e$ the probability of error on future performance when the classifier is trained on the given data set.

4) Denote by $E\{\hat{P}_e\}$ the expected error probability on future performance over the distribution of training sets. $\hat{P}_e$ is an estimate of $E\{\hat{P}_e\}$ and both approach the "actual" error probability P_e as the number of pattern samples approaches infinity. $\hat{P}_e$ and $E\{\hat{P}_e\}$ are also known in the literature as error rates of the sample-based classifier design or decision rule. It is obvious that $\hat{P}_e^B \geq P_e^B$, although it is not necessarily true that $\hat{P}_e^c \geq P_e^c$ unless the underlying distributions are such that the given classifier is Bayes optimal.

5) Denote the "apparent" error probability by P_e(app). For example, P_e^B(app) is obtained by estimating the probability distributions or their parameters and subsequently substituting these estimated values into the expression for error probability. The apparent Bayes error probability is given by

$$P_e^B(\text{app}) = 1 - \int \max_i \{\hat{P}(X/C_i)\hat{P}(C_i)\}\, dX$$

where $\hat{P}(X/C_i)$ and $\hat{P}(C_i)$ are estimates of $P(X/C_i)$ and $P(C_i)$, respectively.

Alternately, one can consider the "apparent" error probability to be that obtained when the sample-based classifier design or decision rule is tested on an infinite number of pattern samples coming from distributions, the parameters of which take on the estimated rather than the true values. P_e^B(app) may be greater or less than P_e^B, but has a tendency to be optimistically biased.

6) For 1), 3), 4), and 5) and any classifier considered in 2), denote by $P_{e_{ij}}$ the *transition probability of error* that a pattern belonging to class i is classified into class j, for i, $j = 1,2,\cdots,M$, $i \neq j$, where there are M classes.

7) For 1), 3), 4), and 5) and any classifier considered in 2), denote by $P_{e|C_i}$ the *class-conditional probability of error*, i.e., the probability that any one pattern belonging to class i is misclassified. It follows that

$$P_{e|C_i} = \sum_{\substack{j=1 \\ j \neq i}}^{M} P_{e_{ij}}.$$

8) For 1), 3), 4), and 5) and any classifier considered in 2), denote by $P_{e|X}$ the *conditional probability of error* given a particular unclassified pattern. It follows that

$$P_e = \int P(X)P_{e|X}\, dX$$

where $P(X)$ is the unconditional, or mixture, distribution.

In most pattern recognition problems one is interested in $\hat{P}_e$. However, $\hat{P}_e$ cannot be obtained exactly because, by definition, all the available pattern samples are used for training the classifier and hence none are left for testing it. Several methods are available for estimating $\hat{P}_e$. Some of these methods are described in the following. Emphasis is placed on nonparametric techniques, since usually nothing is known about the distributions.

There are two basic approaches to the problem of estimation of misclassification: the nonparametric approach, which is almost always used in problems such as character recognition; and the parametric approach, in which it is

Manuscript received March 10, 1973; revised January 22, 1974.
The author is with the School of Computer Science, McGill University, Montreal, P.Q., Canada.

assumed that the unknown distributions belong to a parametric family. The nonparametric methods will be considered first.

Let $\{X,\theta\} = \{X_1, \theta_1; X_2, \theta_2; \cdots; X_N, \theta_N\}$ be the set of N pattern samples available, where X_i and θ_i denote, respectively, the measurement information and the label or classification information of the ith pattern sample. It is assumed that each θ_i associated with X_i is the correct label, i.e., the pattern samples have been correctly preclassified.

Method 1

The first method considered here is the resubstitution, or R method, which consists of the following steps.

1) The classifier is trained on $\{X,\theta\}$.

2) The classifier is tested on $\{X,\theta\}$.

Let the resulting proportion of errors encountered during testing be denoted by $\hat{P}_e[R]$. When pattern recognition as a field of study was still in its infancy, $\hat{P}_e[R]$ was a popular method of estimating $\hat{P}_e$. As a matter of fact, this method of estimating $\hat{P}_e$ was suggested by some statisticians in discriminant analysis long ago [161].

Researchers in the field of pattern recognition soon became interested in the "generalizing" capability of a "learning" machine (adaptive classifier), which gave rise to methods 2)–4). It should be noted that although the three methods are discussed separately for the purpose of clarity and historical perspective, they are all special cases of the method referred to by some statisticians as cross-validation.

Method 2

The second method under investigation is the holdout, or H method, which can be described as follows.

1) Partition $\{X,\theta\}$ into two mutually exclusive sets $\{X,\theta\}_\alpha$ and $\{X,\theta\}_\beta$ such that

$$\{X,\theta\}_\alpha = \{X_1, \theta_1; X_2, \theta_2; \cdots; X_{N(\alpha)}, \theta_{N(\alpha)}\}$$

$$\{X,\theta\}_\beta = \{X_{N(\alpha)+1}, \theta_{N(\alpha)+1}; \cdots; X_N, \theta_N\}$$

and $N(\beta) = N - N(\alpha)$.

2) Train the classifier on $\{X,\theta\}_\alpha$.

3) Test the classifier on $\{X,\theta\}_\beta$.

Let the proportion of errors observed during testing be denoted by $\hat{P}_e[H]$. Traditionally 50 percent of the available samples have been used for training, and 50 percent for testing. The H method was analyzed by Highleyman [89], who indicated a method for obtaining confidence intervals on the results and presented graphs showing how a finite data set of size N should be partitioned between training and test sets for various values of N. However, Kanal and Chandrasekaran [112] showed that Highleyman's analysis and the resulting graphs are valid only when N is sufficiently large, whereas the problem of estimation of misclassification is of most concern when N is small. Additional work on obtaining confidence intervals and partitioning the data set can be found in [16], [17], [50], [51], [118], [127], and [143]. Researchers using the R and H methods soon reported large discrepancies between $\hat{P}_e[R]$ and $\hat{P}_e[H]$. Some of the important works that discuss these discrepancies are [11], [18], [31], [45], [46], [88], and [142]. It was observed that $\Delta\hat{P}_e(H - R) \triangleq \hat{P}_e[H] - \hat{P}_e[R]$ was usually positive, and it was conjectured [45], [88], that the value of $\Delta\hat{P}_e(H - R)$ was inversely proportional to sample size and that

$$\lim_{N\to\infty} \{\Delta\hat{P}_e(H - R)\} = 0.$$

As it turns out, although the R method uses the data efficiently, it is an overly optimistic estimate of performance. Furthermore, unless N is large, the H method tends to give an overly pessimistic estimate of performance, as well as an unreliable one because the value of $\hat{P}_e[H]$ for a given data set depends on the partitioning of $\{X,\theta\}$. The H method also uses the data in an inefficient manner. It can, however, be made more reliable by averaging $\hat{P}_e[H]$ over all possible partitions of fixed size.

1) Partition $\{X,\theta\}$ into K randomly chosen pairs of sets of equal size

$$\{X,\theta\}_\alpha^1: \{X,\theta\}_\beta^1, \cdots, \{X,\theta\}_\alpha^K: \{X,\theta\}_\beta^K$$

such that for $i = 1,2,\cdots,K$, $\{X,\theta\}_\alpha^i$ and $\{X,\theta\}_\beta^i$ are mutually exclusive.

2) For $i = 1,2,\cdots,K$ train the classifier on $\{X,\theta\}_\alpha^i$ and test it on $\{X,\theta\}_\beta^i$, letting the resulting proportion of errors be denoted by $\hat{P}_e[H]_i$.

3) An estimate of the expected value of $\hat{P}_e[H]$ over the partitions is then given by

$$\frac{1}{K}\sum_{i=1}^{K} \hat{P}_e[H]_i. \tag{1}$$

This method of improving the reliability of the estimate was mentioned by Duda and Hart [46], and is known in some circles as "data shuffling" [68]. Although the estimate in (1) uses the data more efficiently than the H method, it still uses only half of the available data for training each time. Furthermore, the final result is still overly pessimistic.

A method which has come to be known in North American circles as the U method or "leave-one-out" method goes a long way towards making efficient use of the data and yielding an estimate of performance with a small amount of bias compared to the previous methods.

Method 3—(U Method)

1) Take one pattern sample (X_i,θ_i) out of $\{X,\theta\}$. Then define

$$\{X,\theta\}_i \triangleq \{X_1, \theta_1; \cdots; X_{i-1}, \theta_{i-1}; X_{i+1}, \theta_{i+1}; \cdots; X_N, \theta_N\}.$$

2) Train the classifier on $\{X,\theta\}_i$.

3) Test the classifier on (X_i,θ_i). If X_i is classified into the category associated with θ_i, set $e_i = 0$; otherwise set $e_i = 1$, where e_i acts as an error indicator.

4) Do steps 1)–3) for $i = 1,2,\cdots,N$ to obtain values for e_i, $i = 1,2,\cdots,N$.

5) The estimate of $\hat{P}_e$, denoted by $\hat{P}_e[U]$, is then computed as follows:

$$\hat{P}_e[U] = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{2}$$

In the statistics literature the U method is attributed to Lachenbruch, [90], [165], who published results on it as

early as 1967 [119]. However, the U method has been under investigation in the pattern recognition literature since the early 1960's. Lunts and Brailovskiy [128] attribute the U method, which they refer to as a "sliding" estimate, to Weinzweig, and they themselves published experimental and theoretical work on it as early as 1964 [21], [22], and [128]. An experimental comparison of the U method with the R method and the actual error probability for different ratios of sample size to feature size (dimensionality) for different nonnormal distributions is given in [124].

In spite of its advantages with regard to bias, the U method suffers from at least two disadvantages. Denote by $E\{\hat{P}_e[U]\}$ the expected value of $\hat{P}_e[U]$ over the distribution of training sets. Although it is desirable to have $E\{\hat{P}_e[U]\}$ "close" to the actual error probability, it is probably more important to use an estimator with a small variance. Hence, an estimator which is more biased than the U method but has a much smaller variance may be preferred by a researcher who may then have more confidence about his particular result on his particular data set. This problem has been considered by Lunts and Brailovskiy [128], who also derive an expression for the variance of $\hat{P}_e[U]$ under certain restrictions. Glick has shown that the U method has much greater variance than the R method for discrete distributions and, in fact, the U method in some sense achieves bias reduction in exactly the "worst" way for the discrete case.[1] Recently Lissack and Fu [126] have proposed a method which they call the F method, and have reported experimental results on Gaussian data. They found that the F method was less biased and had smaller variance than the U method. A second practical disadvantage of the U method is that it requires excessive computation in the distribution-free-case, in the form of N training sessions, unless N is very small. For the case of Gaussian distributions a certain amount of computation can be saved [63], [64], [119]. To combat this disadvantage of the U method the following method, also referred to as the rotation or Π method, was proposed in [171] and [173].

Method 4—(Π *Method*)

1) Take a small subset of pattern samples

$$\{X,\theta\}_i^{TS} \triangleq \{X_1, \theta_1; X_2, \theta_2; \cdots; X_P, \theta_P\}$$

such that $1 \leq P \ll N$ and N/P is an integer, $P/N \leq \frac{1}{2}$. Then

$$\{X,\theta\}_i^{TR} \triangleq \{X_{p+1}, \theta_{p+1}; \cdots; X_N, \theta_N\}.$$

2) Train the classifier on $\{X,\theta\}_i^{TR}$.

3) Test the classifier on $\{X,\theta\}_i^{TS}$ to obtain a proportion of errors denoted by $\hat{P}_e[\Pi]_i$.

4) Do steps 1)–3) for $i = 1,2,\cdots,N/P$ such that $\{X,\theta\}_i^{TS}$ and $\{X,\theta\}_j^{TS}$ are disjoint for $i = 1,2,\cdots,N/P$, $J = 1,2,\cdots,N/P$, and $i \neq j$.

5) The resulting estimate of $\hat{P}_e$ is computed as

$$\frac{P}{N} \sum_{i=1}^{N/P} \hat{P}_e[\Pi]_i. \tag{3}$$

[1] Ned Glick, personal communication.

Note that when $P = 1$ the Π method reduces to the U method. On the other hand, when $P = N/2$ the Π method reduces essentially to the H method, where the roles of training and testing are interchanged. This is the well-known "cross-validation in both directions" method [139].[2] The Π method is also considered in [128]. Obviously, the Π method is a compromise between the U and H methods. One would expect the Π method to be less biased than the H method (depending on the values of P, N, and λ, where λ denotes feature size, dimensionality, or the number of parameters to be estimated) and to require less computation than the U method. Therefore, it is a method well suited to "medium-sized" data sets. Some experimental results on this method of estimating $\hat{P}_e$ are given in [99], [100], [172], and [174]. This method of cross-validation has the flavor of the estimation methods used in statistics to reduce bias that are referred to as "jackknifing" procedures [138].

Two problems closely related to the estimation of the probability of misclassification from a given data set are: 1) reducing the bias of the estimates of the parameters that results when designing the classifier, especially when the training set is small, and 2) estimating the stability of the classifier or estimated parameters based on the given data set. The jackknife [138] serves the dual purpose of eliminating bias in the estimates of the parameters and giving an honest measure of variability, based on the training data itself. For example, consider a linear discriminant function $g(X)$. Let the data be divided into k subgroups and let $g_{\text{all}}(X)$ and $g_j(X)$ denote the discriminant functions computed using the entire data set and using all the data left after omitting the jth subgroup, respectively. The jackknifed discriminant function is then given by

$$g^*(X) = kg_{\text{all}}(X) - \frac{k-1}{k} \sum_{j=1}^{k} g_j(X). \tag{4}$$

In [139] Mosteller and Tukey propose a "leave-two-out" method[3] in which jackknifing and cross-validation are carried out simultaneously. At each step one pattern sample is put aside for cross-validation while another is successively removed from the remaining group of size $N - 1$ in order to obtain a jackknifed classifier design. These methods have been applied to an authorship classification problem in [138]–[141].

Several studies [56], [57] show that these estimates of $\hat{P}_e$ converge to $\hat{P}_e$ as $N \to \infty$. In particular, a quantity such

[2] This method is referred to as "double cross-validation" in the psychology literature [137], [145]. A further extension of these methods is possible, as indicated by Norman [145]. For example, one can use "triple cross-validation" [145] to estimate the probability of misclassification of the best subset of n out of N features for a specified feature-selection criterion. The data set is first partitioned into three subsets. With knowledge of each subset of data, a subset of n features is chosen with the specified feature-selection criterion. For each of the three subsets of n features the classifier is trained with a data subset not used in the feature subset selection procedure. Finally, the classifier incorporating a particular feature subset is tested on the third data subset. The average of the three results is a measure of the performance of the best n features according to a given feature-selection criterion.

[3] It should be kept in mind that this "leave-two-out" method is not the same as the "second-order-jackknife." The latter involves leaving two out for the purpose of bias reduction and has nothing to do with cross-validation. Work on the "second-order-jackknife" can be found in Adams *et al.*, *Ann. Math. Statist.*, vol. 42, pp. 1606–1612, 1971.

as $E\{\hat{P}_e[R]\}$ approaches $\hat{P}_e$ from below whereas another quantity such as $E\{\hat{P}_e[\Pi]\}$ approaches $\hat{P}_e$ from above. Therefore, another estimate for $\hat{P}_e$ can be defined as

$$\hat{P}_e^* = W(N,P,\lambda)\frac{P}{N}\sum_{i=1}^{N/P}\hat{P}_e[\Pi]_i + [1 - W(N,P,\lambda)]\hat{P}_e[R] \tag{5}$$

where $0 \leq W(N,P,\lambda) \leq 1$. In fact, since (3) and $\hat{P}_e[R]$ are estimates of $\hat{P}_e$ biased in opposite directions, it should be possible to determine the function $W(N,P,\lambda)$, at least empirically, such that $\hat{P}_e^*$ is an essentially unbiased estimate of $\hat{P}_e$. For example, in the work of Foley [56], the average of the results on the test set and training set provides quite a good estimate of the actual error probability, even for small N. In [174], $\hat{P}_e^*$ was applied to a problem in medical diagnosis and it was found empirically that for $W(N,P,\lambda) =$ constant $= 1/2$ and $P/N = 1/10$, where $N = 300$, $\hat{P}_e^*$ was essentially equal to $\hat{P}_e[U]$. Furthermore, a tremendous saving was realized, because 300 training sessions were needed to obtain $\hat{P}_e[U]$, whereas for $\hat{P}_e^*$ only 11 training sessions were needed—one to obtain $\hat{P}_e[R]$ and ten to obtain (3).

Studying the error probability on the training and testing sets as a function of N, the number of pattern samples, is not the whole story. The estimation of $\hat{P}_e$ is also intimately related to the number of features or measurements used by the classifier.[4] Early experimental observations of this dependence in pattern recognition, discriminant analysis, and disease diagnosis can be found in [4], [5], [33], [38], [49], [67], [73], and [175]. Some theoretical work along the same lines is given in [1], [28], [29], [46], [56], [57], [96], [98], [111], and [154]. The problem of estimating $\hat{P}_e$ is further complicated by the fact that it depends on whether there exist dependencies among the features and ultimately on the actual distributions for the problem at hand. One measure of error probability not yet defined here is the "problem-average" error rate or, as defined by Hughes [96], the "mean-recognition-accuracy" denoted by $\bar{P}_e(L,N)$, where N is the number of samples and L is the number of "cells" or values X can take in the discrete case. $\bar{P}_e(L,N)$ represents the error probability averaged over all possible pattern recognition problems or distributions on X. Further results on $\bar{P}_e(L,N)$ are given in [1], [28], [29], [98], and [111].

The literature contains a wide variety of other methods (the parametric methods) for estimating various measures of probability of misclassification when certain *a priori* information concerning the distributions is available. These methods usually invoke the normality assumption. An example is the D method in [123], where it is assumed that there are two classes ($M = 2$) with means μ_1 and μ_2, and equal covariance matrix Σ. Obviously, if the parameters were known there would be no problem and the class-conditional Bayes probabilities of misclassification would be given by

$$P_{e|C_1}^B = P_{e|C_2}^B = \Phi(-\delta/2)$$

where $\delta^2 = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$ is the Mahalanobis distance, Φ is the cumulative normal distribution, and T denotes the transpose. When the parameters are not known the D method yields an estimate of $P_{e|C_i}^B$, $i = 1,2$, which is given by

$$P_i = \Phi(-D/2)$$

where D^2 is the Mahalanobis sample distance when the sample means $\bar{X}_1$ and $\bar{X}_2$ and the sample covariance matrix S are substituted for μ_1, μ_2, and Σ, respectively. Various modifications of this D method exist, in which different types of estimates of the Mahalanobis distance are used. Discussions and comparisons of all these methods can be found in [23], [119], [120], [122], [123], [151], [162]–[165], and [167]. It should be noted that the D method yields the "apparent" error rate discussed earlier for the case of normal distributions.

The parametric approach to estimation of misclassification involves two other theoretical aspects of the problem: the distributions of classification statistics on one hand [19], [20], [84], [102]–[110], [133]–[135], [146], [158], [159], and the convergence properties of the estimators on the other [54], [55], [74]–[76], [176]. For example, in a typical approach, for the two-category problem with Gaussian distributions and known equal covariance matrices, John [102] proposes a classification procedure based on the calculation of a classification statistic and derives an analytic expression, in the form of an infinite series, for the probability of misclassification as a function of N_1 and N_2, where $N_1 + N_2 = N$ and N_i is the number of training pattern samples in the ith category. Furthermore, for the case of M categories, $M > 2$, he derives an upper bound on the probability of misclassification. In order to set up the classification procedure and to study performance characteristics, such as the probability of misclassification, it is necessary to know the distribution of the statistic used in the classification procedure. In [104] John gives the exact distribution of several classification statistics. Glick [75], [76] extensively covers what he calls "plug-in" estimates (for Gaussian distributions, actually the D method of Lachenbruch and Mickey [123]) and "deletion-counting" estimates (actually the "sliding" estimates in [128] or the U method in [123]). He also considers these estimates for various decision rules such as the nearest-neighbor rule.

The "deleted nearest-neighbor" estimate of $\hat{P}_e$ was first proposed by Cover [39]. Since then further theoretical work has been done on the estimate and on an interesting modification of it in [178], [179], and [187]. The modification is essentially a recursively updated nearest-neighbor estimate as new pattern samples are added. Experimental work with the deleted nearest-neighbor estimate can be found in [174] and [184]. Additional work on the convergence of nearest-neighbor estimates is given in [41], [42], [78], [82], [85], [147]–[150]. Estimating error probabilities using nearest-neighbor rules requires a great

[4] Anderson and Isenhour [6] have recently done extensive Monte Carlo studies on dimensionality λ versus sample size N as related to linear separability. They found that even for constant N/λ the probability of error on the testing set tends to decrease as λ increases for values of $N/\lambda < 3$.

deal of storage when data sets are large. Hence some attempts have been made at decreasing the size of the training data; in some cases better estimates of $\hat{P}_e^B$ are obtained than those obtained by using the entire training set [53], [83], [87], [147], [170], and [186].[5]

Some further miscellaneous work concerning the estimation of $\hat{P}_e$ and related problems can be found in [3], [12], [13], [15], [34], [36], [37], [58], [81], [116], [125], [152], [156], and [166]. Friedman defines $\Delta = \hat{P}_e^B - P_e^B$ as the "error determined by a classifier," the parameters of which are estimated from a particular training set. For the two-class problem with equal *a priori* probabilities, he derives expected values of Δ and examines their asymptotic behavior for a few simple univariate Gaussian cases. Wilkins and Ford [185] discuss the effects of unrepresentative samples present in the training sets.

The reader should be reminded that estimating the error on a given data set by making proper use of the best procedures available is only half the problem. One may have a large enough data set relative to the number of parameters to be estimated, but does it adequately represent the variability in the data for the problem one is trying to solve? Hence a problem of perhaps even greater importance than the previous one is the adequacy of the data set itself. For example, in speech recognition and alphanumeric handprinted character recognition, thousands of samples are required to represent the true variability in the data from the population at large. This is the price which must be paid when the underlying class-conditional distributions are not known.

Acknowledgment

The author is grateful to the referees for their helpful comments and to T. Cover for reading the manuscript and suggesting stylistic changes, as well as removing ambiguities. The references on cross-validation in the psychology literature were compiled by L. Goldberg and brought to my attention by E. Mark Gold, who also contributed to stimulating discussions on the jackknife and cross-validation methods. Thanks are due to all those who kindly sent in reprints of their work. The author would be grateful to be informed of any additional references on these topics that readers may be aware of and that have been left out of this bibliography.

References

[1] K. Abend, T. J. Harley, Jr., B. Chandrasekaran, and C. F. Hughes, "Comments 'On the mean accuracy of statistical pattern recognizers'," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-15, pp. 420–423, May 1969.

[2] A. A. Afifi and S. P. Azen, *Statistical Analysis: A Computer Oriented Approach.* New York: Academic Press, 1972, pp. 244–245.

[3] R. Albreht and W. Werner, "Error analysis of a statistical decision method," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 34–38, Jan. 1964.

[4] D. C. Allais, "Selection of measurements for prediction," Stanford Electron. Res. Lab., Stanford, Calif., Rep. SEL-64-115, TR 6103-9, Nov. 1964.

[5] ——, "The problem of too many measurements in pattern recognition and prediction," in *IEEE Int. Conv. Rec.*, vol. 14, pt. 2, 1966, pp. 124–130.

[6] D. N. Anderson and T. L. Isenhour, "Empirical studies of separability and prediction using threshold logic units," *Pattern Recognition*, vol. 5, pp. 249–258, Sept. 1973.

[7] T. W. Anderson, "Classification by multivariate analysis," *Psychometrika*, vol. 16, pp. 31–50, 1951.

[8] ——, "Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 17–35.

[9] ——, "An asymptotic expansion of the distribution of the studentized classification statistic," *Ann. Statist.*, vol. 1, pp. 964–972, Sept. 1973.

[10] T. W. Anderson, S. D. Gupta, and G. P. Styan, *A Bibliography of Multivariate Statistical Analysis.* New York: Wiley, 1972.

[11] P. Armitage, "Recent developments in medical statistics," *Rev. Int. Stat. Inst.*, vol. 34, pp. 27–42, 1966.

[12] S. P. Azen and A. A. Afifi, "Asymptotic and small-sample behaviour of estimated Bayes rules for classifying time dependent observations," *Biometrics*, vol. 28, pp. 898–998, Dec. 1972.

[13] ——, "Two models for assessing prognosis on the basis of successive observations," *Mathematical Biosciences*, vol. 14, pp. 169–176, 1972.

[14] M. S. Bartlett, "An inverse matrix adjustment arising in discriminant analysis," *Ann. Math. Statist.*, vol. 22, pp. 107–111, 1951.

[15] M. S. Bartlett and N. W. Please, "Discrimination in the case of zero mean differences," *Biometrika*, vol. 50, pp. 17–21, 1963.

[16] L. S. Belobragina and V. K. Eliseev, "Statistical investigation of a recognition algorithm for type-written symbols," *Kibernetika*, vol. 3, no. 6, pp. 65–72, 1967; see also English translation in *Cybernetics*, vol. 3, no. 6, pp. 51–57, 1967.

[17] ——, "Statistical estimation of recognition error probability from experimental data," *Kibernetika*, vol. 3, no. 4, pp. 81–89, 1967; see also English translation in *Cybernetics*, vol. 3, no. 4, pp. 67–73, 1967.

[18] W. W. Bledsoe, "Further results on the N-tuple pattern recognition method," *IRE Trans. Electron. Comput.* (Corresp.), vol. EC-10, p. 96, Mar. 1961.

[19] A. H. Bowker, "A representation of Hotelling's T^2 and Anderson's statistic W in terms of simple statistics," *Contributions to Probability and Statistics.* Stanford, Calif.: Stanford Univ. Press, 1960, pp. 142–150.

[20] A. H. Bowker and R. Sitgreaves, "Asymptotic expansion for the distribution function of the W-classification statistic," in *Studies in Item Analysis and Prediction*, H. Solomon, Ed. Stanford, Calif.: Stanford Univ. Press, 1961.

[21] V. L. Brailovskiy, "An object recognition algorithm with many parameters and its applications," *Eng. Cybern.* (USSR), no. 2, pp. 22–30, 1964.

[22] V. L. Brailovskiy and A. L. Lunts, "The multiparameter recognition problem and its solution," *Eng. Cybern. (USSR)*, no. 1, pp. 13–22, 1964.

[23] J. D. Broffitt, "Estimating the probability of misclassification based on discriminant function techniques," Ph.D. dissertation, Colorado State Univ., Fort Collins, 1969.

[24] ——, "Minimum variance estimators for misclassification probabilities in discriminant analysis," *J. Multivariate Analysis*, vol. 3, pp. 311–327, Sept. 1973.

[25] J. D. Broffitt and J. S. Williams, "Distributions of functions of $A = (ZZ')^{-1/2}Z$ where the density of Z satisfies certain symmetry conditions," Dep. Statistics, Univ. Iowa, Iowa City, Tech. Rep. 7, 1971.

[26] L. S. Chan, "The treatment of missing values in discriminant analysis," M.S. thesis, Univ. California, Los Angeles, 1970.

[27] L. S. Chan and O. J. Dunn, "The treatment of missing values in discriminant analysis—1. The sampling experiment," *J.A.S.A.*, vol. 67, no. 338, 473–477, June 1972.

[28] B. Chandrasekaran, "Independence of measurements and the mean recognition accuracy," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 452–456, July 1971.

[29] B. Chandrasekaran and A. K. Jain, "Quantization of independent measurements and recognition performance," presented at the 1972 IEEE Int. Symp. Information Theory.

[30] Z. Chen and K. S. Fu, "Nonparametric Bayes risk estimation for pattern classification," in *Proc. IEEE Conf. on Systems, Man, and Cybernetics*, Boston, Mass., Nov. 5–7, 1973.

[31] C. K. Chow, "A recognition method using neighbour dependence," *IRE Trans. Electron. Comput.*, vol. EC-11, pp. 683–690, Oct. 1962.

[32] ——, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 41–46, Jan. 1970.

[33] J. T. Chu and J. C. Chueh, "Error probabilities in decision functions for character recognition," *J. Ass. Comput. Mach.*, vol. 14, pp. 273–280, Apr. 1967.

[34] W. G. Cochran, "On the performance of linear discriminant functions," *Bull. Int. Statist. Inst.*, vol. 34, pp. 435–447, 1961;

[5] Applications of these methods to judicial decisions can be found in [129].

also in *Technometrics*, vol. 6, pp. 179–190, 1964.
[35] ——, "Commentary on estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 204–205, Feb. 1968.
[36] W. G. Cochran and C. E. Hopkins, "Some classification problems with multivariate qualitative data," *Biometrics*, vol. 17, pp. 10–32, 1961.
[37] J. Cornfield, "Joint dependence of risk on coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis," *Fed. Proc.* (Abstr.), vol. 21, no. II, pp. 58–61, 1962.
[38] ——, "Discriminant functions," *Rev. Int. Statist. Inst.*, vol. 35, pp. 142–153, 1967.
[39] T. M. Cover, "Learning in pattern recognition," *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic Press, 1969, pp. 111–132.
[40] ——, "Geometrical and statistical properties of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, pp. 326–334, June 1965.
[41] ——, "Rates of convergence of nearest neighbor decision procedures," presented at 1st Annu. Hawaii Conf. Systems Theory, Jan. 1968, pp. 413–415.
[42] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.
[43] E. E. Cureton, "Approximate linear restraints and best predictor weights," *Educational and Psychological Measurement*, vol. 11, pp. 12–15, 1951.
[44] ——, "Validity, reliability, and boloney," *Educational and Psychological Measurement*, vol. 10, pp. 94–96, 1950.
[45] R. O. Duda and H. Fossum, "Pattern classification by iteratively determined linear and piecewise linear discriminant functions," *IEEE Trans. Electron. Comput.*, vol. EC-15, pp. 220–232, Apr. 1966.
[46] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[47] O. J. Dunn, "Some expected values for probabilities of correct classification in discriminant analysis," *Technometrics*, vol. 13, pp. 345–353, May 1971.
[48] O. J. Dunn and P. D. Varady, "Probabilities of correct classification in discriminant analysis," *Biometrics*, vol. 22, pp. 908–924, 1966.
[49] S. E. Estes, "Measurement selection for linear discriminants used in pattern classification," Ph.D. dissertation, Stanford University, Stanford, Calif., 1965.
[50] V. K. Eliseev, "Statistical methods of the experimental investigation of reliability (quality) of recognition systems," in *Proc. Seminar Pattern Recognition and Construction of Reading Automata* (in Russian), no. 1, Kiev, 1966.
[51] ——, "Statistical investigation of the reliability of a reading automaton with optical correlation," collection in *Reading Automata* (in Russian), izd-vo Naukova dumka, Kiev, 1965.
[52] T. B. Farver, "Stepwise selection of variables in discriminant analysis," Ph.D. dissertation, Univ. California, Los Angeles, 1972.
[53] F. P. Fischer, "A preprocessing algorithm for nearest neighbour decision rules," in *Proc. Nat. Electron. Conf.*, Chicago, Ill., Dec. 7–9, 1970, pp. 481–485.
[54] E. Fix and J. L. Hodges, "Discriminatory analysis: nonparametric discrimination: small sample performance," U.S. School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rep. 11, AF 4(128)-31, 1952.
[55] ——, "Nonparametric discrimination: consistency properties," U.S. School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rep. 4, AF 41(128)-31, 1951.
[56] D. H. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 618–626, Sept. 1972.
[57] ——, "The probability of error on the design set as a function of the sample size and feature size," Ph.D. dissertation, Syracuse Univ., Syracuse, N.Y., June 1971; also Tech. Rep. RADC-TR-71-171.
[58] I. Francis, "Inference in the classification problem," Ph.D. dissertation, Harvard Univ., Cambridge, Mass., 1966.
[59] H. D. Friedman, "On the expected error in the probability of misclassification," *Proc. IEEE*, vol. 53, pp. 658–659, June 1965.
[60] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1972, ch. 5.
[61] K. Fukunaga and D. Kessell, "Nonparametric Bayes error estimation using unclassified samples," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 434–439, July 1973.
[62] ——, "Application of optimum error-reject functions," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-18, pp. 814–817, Nov. 1972.
[63] ——, "Error evaluation and model validation in statistical pattern recognition," Sch. Elec. Eng., Purdue Univ., Tech. Rep. TR-EE 72-23, Aug. 1972.
[64] ——, "Estimation of classification error," *IEEE Trans. Comput.*, vol. C-20, pp. 1521–1527, Dec. 1971.
[65] ——, "A model for error estimation," in *Proc. 1971 IEEE Conf. Decision and Control* (including *10th Symp. Adaptive Processes*), Miami Beach, Fla., Dec. 15–17, pp. 351–352.
[66] K. Fukunaga and T. F. Krile, "Calculation of Bayes recognition error for two multivariate Gaussian distributions," *IEEE Trans. Comput.*, vol. C-18, pp. 220–229, Mar. 1969.
[67] W. R. Gaffey, "Discriminatory analysis: perfect discrimination as the number of variables increases," U.S. School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Feb. 1951.
[68] M. F. Gardiner *et al.*, "The interpretation of statistical measures from discriminant analysis in evoked response experiments," in *Proc. 5th Hawaii Int. Conf. Systems Sciences, Computers in Biomedicine*, 1972, pp. 235–237.
[69] S. Geisser, "Predictive discrimination," in *Proc. Int. Symp. Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic Press, 1966, pp. 149–163.
[70] ——, "Posterior odds for multivariate normal classification," *J. Roy. Stat. Soc.*, series B, vol. 26, pp. 69–76, 1964.
[71] ——, "Estimation associated with linear discriminants," *Ann. Math. Statist.*, vol. 38, pp. 807–817, 1967.
[72] S. G. Ghurye and I. Olkin, "Unbiased estimation of some multivariate probability densities and related functions," *Ann. Math. Statist.*, vol. 40, no. 4, pp. 1261–1271, 1969.
[73] E. S. Gilbert, "On discrimination using qualitative variables," *J. Amer. Statist. Ass.*, vol. 63, pp. 1399–1412, Dec. 1968.
[74] N. Glick, "Sample-based classification procedures derived from density estimators," *J.A.S.A.*, vol. 67, pp. 116–122, Mar. 1972.
[75] ——, "Sample-based multinomial classification," *Biometrics*, vol. 29, pp. 241–256, June 1973.
[76] ——, "Classification procedures based on random samples from a mixture," Univ. Chicago, Chicago, Ill., Tech. Rep., 69 pp. (obtainable from author, Dep. Mathematics, Univ. British Columbia, Vancouver, B.C., Canada).
[77] E. Mark Gold, "Simple formulae for jackknife crossvalidation," July 1972, informal report available from the author, Department d'Informatique, Univ., Montreal, Montreal, P.Q., Canada.
[78] M. Goldstein, "k_n-Nearest neighbour classification," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 627–630, Sept. 1972.
[79] H. L. Gray and W. R. Schucany, *The Generalized Jackknife Statistic*. New York: Marcel Dekker, Inc., 1973.
[80] S. D. Gupta, "Theories and methods in classification: a review," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 77–137.
[81] T. J. Harley, L. N. Kanal, and N. C. Randall, "System considerations for automatic imagery screening," in *Pictorial Pattern Recognition*, G. C. Cheng *et al.*, Eds. Washington, D.C.: Thompson, 1968, pp. 15–31.
[82] P. E. Hart, "An asymptotic analysis of the nearest neighbour decision rule," Stanford Electronics Labs., Stanford, Calif., Tech. Rep. 1828-2, May 1966.
[83] ——, "The condensed nearest neighbour rule," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-14, pp. 515–516, May 1968.
[84] H. L. Harter, "On the distribution of Wald's classification statistic," *Ann. Math. Statist.*, vol. 22, pp. 58–67, 1951.
[85] M. E. Hellman, "The nearest neighbour classification rule with a reject option," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-6, pp. 179–185, July 1970.
[86] P. Herzberg, "The parameters of cross-validation," *Psychometrika* (Suppl.), vol. 34, no. 16, 1969.
[87] R. P. Heydorn, "Nonparametric classification," Ph.D. dissertation, Ohio State Univ., Columbus, Ohio, 1971.
[88] W. H. Highleyman, "Linear decision function, with application to pattern recognition," *Proc. IRE*, vol. 50, pp. 1501–1514, June 1962.
[89] ——, "The design and analysis of pattern recognition experiments," *Bell Syst. Tech. J.*, vol. 41, pp. 723–744, Mar. 1962.
[90] M. Hills, "Allocation rules and their error rates," *J. Roy. Statist. Soc.*, series B, no. 28, pp. 1–31, 1966.
[91] ——, "Discrimination and allocation with discrete data," *J. Roy. Statist. Soc.*, series C, vol. 16, no. 2, pp. 237–250, 1967.
[92] P. G. Hoel and R. P. Peterson, "A solution to the problem of optimum allocation," *Ann. Math. Statist.*, vol. 20, pp. 433–438, 1949.
[93] H. Hudimoto, "On the distribution-free classification of an individual into one of two groups," *Ann. Inst. Stat. Math.*, vol. 8, pp. 105–112, 1956.
[94] ——, "A note on the probability of the correct classification when the distributions are not specified," *Ann. Inst. Stat. Math.*, vol. 9, pp. 31–36, 1958.
[95] ——, "On a distribution-free two-way classification," *Ann. Inst. Stat. Math.*, vol. 16, pp. 247–253, 1964.
[96] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, Jan. 1968.
[97] G. F. Hughes and J. A. Lebo, "Data reduction using information theoretic techniques," Rome Air Development Center, Griffiss Air Force Base, Rome, N.Y., RADC Rep. TR-67-67, Mar. 1967, pp. 45–46.
[98] G. F. Hughes, "Number of pattern classifier design samples per

class," *IEEE Trans. Inform. Theory* (Corresp.), pp. 615–618, Sept. 1969.
[99] A. B. S. Hussain, G. T. Toussaint, and R. W. Donaldson, "Results obtained using a simple character recognition procedure on Munson's handprinted data," *IEEE Trans. Comput.* (Short Notes), vol. C-20, pp. 201–205, Feb. 1972.
[100] A. B. S. Hussain, "Sequential decision schemes for statistical pattern recognition problems with dependent and independent hypotheses," Ph.D. dissertation, Univ. British Columbia, Vancouver, B.C., Canada, June 1972.
[101] S. John, "Errors in discrimination," *Ann. Math. Statist.*, vol. 32, pp. 1125–1144, 1961.
[102] ——, "On some classification problems—I," *Sankhya: Ind. J. Stat.*, vol. 22, pp. 301–308, 1960.
[103] ——, "On classification by the statistics R and Z," *Ann. Inst. Stat. Math.*, vol. 14, pp. 237–246, 1962.
[104] ——, "On some classification statistics," *Sankhya: Ind. J. Stat.*, vol. 22, pp. 309–316, 1960.
[105] ——, "The distribution of Wald's classification statistic when the dispersion matrix is known," *Sankhya: Ind. J. Stat.*, vol. 21, pp. 371–376, 1959.
[106] ——, "Corrigenda to a paper 'On some classification problems—I'," *Sankhya: Ind. J. Stat.*, vol. 23, ser. A, p. 308, 1961.
[107] ——, "Corrigenda to a paper 'On some classification statistics'," *Sankhya: Ind. J. Stat.*, vol. 23, ser. A, p. 308, 1961.
[108] ——, "Further results on classification by W," *Sankhya: Ind. J. Stat.*, vol. 26, ser. A, pp. 39–46, 1964.
[109] ——, "Corrections to 'On classification by the statistics R and Z'," *Ann. Inst. Stat. Math.*, vol. 17, p. 113, 1965.
[110] D. G. Kabe, "Some results on the distribution of two random matrices used in classification procedures," *Ann. Math. Statist.*, vol. 34, p. 181, 1964 (a correction to this article appeared in *Ann. Math. Statist.*, vol. 34, p. 924, 1964).
[111] R. Y. Kain, "The mean accuracy of pattern recognizers with many pattern classes," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-15, pp. 424–425, May 1969.
[112] L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," in *Proc. Nat. Electron. Conf.*, vol. 24, pp. 2–7, 1968; also in *Pattern Recognition*, vol. 3, pp. 225–234, Oct. 1971.
[113] L. Kanal and N. C. Randall, "Recognition system design by statistical analysis," in *Proc. 19th Nat. Conf. Ass. Computing Machinery*, pp. D2-5-1–D2-5-10, 1964.
[114] R. A. Katzell, "Cross-validation of item analyses," *Educational and Psychological Measurement*, vol. 11, pp. 16–22, 1951.
[115] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, vol. 3. London: Griffin, 1966, pp. 328–329.
[116] C. F. Kossack, "A handbook of statistical classification techniques," Purdue Univ., Lafayette, Ind., Res. Rep. 601-866, 1964.
[117] A. Kudo, "The classificatory problem viewed as a two-decision problem," *Memoirs Faculty of Science*, Kyushu University, series A, vol. 13, pp. 96–125, 1959.
[118] E. F. Kushner, "Experiments on the application of the correlation method for the recognition of figures and letters written by hand," in *Proc. Seminar Pattern Recognition and Construction of Reading Automata* (in Russian), Kiev, 1966.
[119] P. A. Lachenbruch, "An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis," *Biometrics*, vol. 23, pp. 639–645, Dec. 1967.
[120] ——, "On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient," *Biometrics*, vol. 24, pp. 823–834, Dec. 1968.
[121] ——, "Some results on the multiple group discriminant problem," in *Discriminant Analysis and Applications*, T.Cacoullos, Ed. New York: Academic Press, 1973, pp. 193–211.
[122] ——, "Estimation of error rates in discriminant analysis," Ph.D. dissertation, Univ. California, Los Angeles, 1965.
[123] P. A. Lachenbruch and R. M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.
[124] P. A. Lachenbruch, C. Sneeringer, and L. T. Revo, "Robustness of the linear and quadratic discriminant functions to certain types of non-normality," *Communications in Statistics*, vol. 1, no. 1, pp. 39–56, 1972.
[125] L. E. Larsen, *et al.*, "On the problem of bias in error rate estimation for discriminant analysis," *Pattern Recognition*, vol. 3, pp. 217–224, Oct. 1971.
[126] T. Lissack and K. S. Fu, "A separability measure for feature selection and error estimation in pattern recognition," Sch. Elec. Eng., Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE-72-15, May 1972.
[127] V. I. Loginov, "Probability estimates of the quality of a pattern recognition decision rule," *Eng. Cybern.* (USSR), no. 6, pp. 112–117, 1968.
[128] A. L. Lunts and V. L. Brailovskiy, "Evaluation of attributes obtained in statistical decision rules," *Eng. Cybern.* (USSR), no. 3, pp. 98–109, 1967.
[129] E. Mackaay and P. Robillard, "Predicting judicial decisions: the nearest neighbour rule and visual representation of case patterns," Publication no. 131, Department d'Informatique, Univ. Montreal, Montreal, P.Q., Canada.
[130] S. Marks, "Discriminant functions when covariance matrices are unequal," Ph.D. dissertation, Univ. California, Los Angeles, 1970.
[131] D. C. Martin and R. R. Bradley, "Probability models, estimation, and classification for multivariate dichotomous populations," *Biometrics*, vol. 28, pp. 203–222, 1972.
[132] G. J. McLachlan, "An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function," *Aust. J. Statist.*, vol. 14, pp. 68–72, 1972.
[133] A. Z. Memon, "Z statistic in discriminant analysis," Ph.D. dissertation, Iowa State Univ., Ames, Iowa, 1968.
[134] A. Z. Memon and M. Okamoto, "The classification statistic W^* in covariate discriminant analysis," *Ann. Math. Statist.*, vol. 41, pp. 1491–1499, 1970.
[135] ——, "Assymptotic expansion of the distribution of the Z statistic in discriminant analysis," *J. Multivariate Analysis*, vol. 1, pp. 294–307, Sept. 1971.
[136] J. Michaelis, "Simulation experiments with multiple group linear and quadratic discriminant analysis," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 225–238.
[137] C. I. Mosier, "Problems and designs of cross-validation," *Educational and Psychological Measurement*, vol. 11, pp. 5–11, 1951.
[138] F. Mosteller, "The jackknife," *Rev. Int. Statist. Inst.*, vol. 39, no. 3, pp. 363–368, 1971.
[139] F. Mosteller and J. W. Tukey, "Data analysis, including statistics," in *Revised Handbook of Social Psychology*, vol. 2, G. Lindzey and E. Aronson, Eds. Reading, Mass.: Addison-Wesley, 1968, ch. 10, pp. 80–203.
[140] F. Mosteller and D. L. Wallace, "Inference in an authorship problem," *J.A.S.A.*, vol. 58, pp. 275–309, 1963.
[141] ——, *Inference in an Authorship Problem: The Federalist*. Reading, Mass.: Addison-Wesley, 1964.
[142] J. H. Munson, R. O. Duda, and P. E. Hart, "Experiments with Highleyman's data," *IEEE Trans. Comput.* (Short Notes), pp. 399–401, Apr. 1968.
[143] R. A. Nashlyunas and A. V. Lashas, "The problem of estimating the reliability of recognition algorithms," collection in *Automatics and Computational Technology* (in Russian), izd-vo Mintis, Vil'nyus, 1965.
[144] G. E. Nicholson, Jr., "Prediction in future samples," in *Contributions to Probability and Statistics*, I. Olkin *et al.*, Eds. Stanford, Calif.: Stanford Univ. Press, 1960, pp. 322–330.
[145] W. T. Norman, "Double-split cross-validation: an extension of Mosier's design, two undesirable alternatives, and some enigmatic results," *J. Appl. Psychol.*, vol. 49, pp. 348–357, 1965.
[146] M. Okamoto, "An asymptotic expansion for the distribution of the linear discriminant function," *Ann. Math. Statist.*, vol. 34, pp. 1286–1301, 1963. Correction in *Ann. Math. Statist.*, vol. 39, p. 1358, 1968.
[147] E. A. Patrick, *Fundamentals of Pattern Recognition*. Englewood Cliffs, N.J.: Prentice-Hall, 1972, ch. 4.
[148] E. A. Patrick and F. P. Fisher, "Generalized k nearest neighbour decision rule," *Inform. Contr.*, vol. 16, pp. 128–152, Apr. 1970.
[149] C. R. Pelto, "Adaptive nonparametric classification," *Technometrics*, vol. 11, pp. 775–792, Nov. 1969.
[150] D. W. Peterson, "Some convergence properties of a nearest neighbour decision rule," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 26–31, Jan. 1970.
[151] R. E. Pogue, "Some investigations of multivariate discrimination procedures, with applications to diagnosis clinical electrocardiography," Ph.D. dissertation, Univ. Minnesota, Minneapolis, 1966.
[152] C. R. Rao, "On a general theory of discrimination when the information on alternate hypotheses is based on samples," *Ann. Math. Statist.*, vol. 25, pp. 651–670, 1954.
[153] P. R. Rayment, "The identification problem for a mixture of observations from two normal populations," *Technometrics*, vol. 14, pp. 911–918, Nov. 1972.
[154] J. Sammon, D. H. Foley, and A. Proctor, "Considerations of dimensionality versus sample size," in *Proc. 1970 IEEE Symp. Adaptive Processes*, Dec. 7–9, Austin, Tex.
[155] W. Schaafsma, "Classifying when populations are estimated," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 339–364.
[156] N. Sedransk, "Contributions to discriminant analysis," Ph.D. dissertation, Iowa State University, Ames, 1969.
[157] N. Sedransk and M. Okamoto, "Estimation of the probabilities of misclassification for a linear discriminant function in the univariate normal case," *Ann. Inst. Stat. Math.*, vol. 23, no. 3, pp. 419–435, 1971.
[158] R. Sitgreaves, "On the distribution of two random matrices

used in classification procedures," *Ann. Math. Statist.*, vol. 23, pp. 263–270, 1952.
[159] ——, "Some results on the distribution of the *W*-classification statistic," in *Studies in Item Analysis and Prediction*. Stanford, Calif.: Stanford Univ. Press, 1961, pp. 241–261.
[160] ——, "Some operating characteristics of linear discriminant functions," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 365–374.
[161] C. A. B. Smith, "Some examples of discrimination," *Ann. Eugen.*, vol. 18, p. 272, 1947.
[162] M. Sorum, "Estimating the expected and the optimal probabilities of misclassification," *Technometrics*, vol. 14, pp. 935–943, Nov. 1972.
[163] ——, "Estimating the probability of misclassification," Ph.D. dissertation, Univ. Minnesota, Minneapolis, 1968.
[164] ——, "Three probabilties of misclassification," *Technometrics*, vol. 14, pp. 309–316, May 1972.
[165] ——, "Estimating the conditional probability of misclassification," *Technometrics*, vol. 13, pp. 333–343, May 1971.
[166] ——, "Estimating the expected probability of misclassification for a rule based on the linear discriminant function: univariate normal case," *Technometrics*, vol. 15, pp. 329–340, May 1973.
[167] M. Sorum and R. J. Buehler, "Some conditional estimation problems with applications to estimating the probability of misclassification," Statist. Dep., Univ. Minnesota, Minneapolis, Tech. Rep. 112, 1968.
[168] M. S. Srivastava, "Evaluation of misclassification errors," *Can. J. Statist.*, vol. 1, no. 1, pp. 35–50, 1973.
[169] D. Stoller, "Univariate two-population distribution-free discrimination," *J.A.S.A.*, vol. 49, pp. 770–777, 1954.
[170] C. W. Swonger, "Sample set condensation for a condensed nearest neighbour decision rule for pattern recognition," in *Frontiers of Pattern Recognition*, Satosi Watanabe, Ed. New York: Academic Press, 1972, pp. 511–526.
[171] G. T. Toussaint, "Machine recognition of independent and contextually constrained contour-traced handprinted characters," M.A.Sc. thesis, Univ. British Columbia, Vancouver, Canada, Dec. 1969.
[172] ——, "Feature evaluation criteria and contextual decoding algorithms in statistical pattern recognition," Ph.D. dissertation, Univ. British Columbia, Vancouver, Canada, Oct. 1972.
[173] G. T. Toussaint and R. W. Donaldson, "Algorithms for recognizing contour-traced handprinted characters," *IEEE Trans. Comput.*, vol. C-19, pp. 541–546, June 1970.
[174] G. T. Toussaint and P. M. Sharpe, "An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis," *Computers in Biology and Medicine*, to be published.
[175] J. R. Ullman, "Experiments with the *n*-tuple method of pattern recognition," *IEEE Trans. Comput.*, vol. C-18, pp. 1135–1137, Dec. 1969.
[176] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya: Ind. J. Stat.*, vol. A28, pts. 2 and 3, pp. 261–270, 1966.
[177] ——, "Nonparametric Bayesian decision procedures for (pattern) classification with stochastic learning," in *Trans. 4th Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*. Prague: Czechoslovak Academy of Sciences, 1967.
[178] T. J. Wagner, "Convergence of the nearest neighbour rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566–571, Sept. 1971.
[179] ——, "Deleted estimates of the Bayes risk," *Ann. Statist.*, vol. 1, pp. 359–362, Mar. 1973.
[180] A. Wald, "On a statistical problem in the classification of an individual into one of two groups," *Ann. Math. Statist.*, vol. 15, pp. 145–162, 1944.
[181] W. G. Wee, "On computational aspects and properties of a pattern classifier," presented at 5th Hawaii Int. Conf. System Sciences, Jan. 1972, pp. 598–600.
[182] R. J. Wherry, "Comparison of cross-validation with statistical inference of betas and multiple *R* from a single sample," *Educational and Psychological Measurement*, vol. 11, pp. 23–28, 1951.
[183] J. Wiggin, *Personality and Prediction*. Reading, Mass.: Addison-Wesley, 1973, pp. 46–49.
[184] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Trans. Comput.*, vol. C-20, pp. 1100–1103, Sept. 1971.
[185] B. R. Wilkins and N. L. Ford, "The analysis of training sets for adaptive pattern classifiers," in *Proc. Conf. Machine Perception of Patterns and Pictures*, Teddington, England, London, England, Apr. 12–14, 1972, pp. 267–275.
[186] D. L. Wilson, "Asymptotic properties of nearest neighbour rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, pp. 408–420, July 1972.
[187] C. T. Wolverton, "Strong consistency of an estimate of the asymptotic error probability of the nearest neighbour rule," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 119–120, Jan. 1973.
[188] M. Zielezny, "A two-stage classification rule based on expected cost," Ph.D. dissertation, Univ. California, Los Angeles, 1971.

Optimum Transmitting Filter in Digital PAM Systems with a Viterbi Detector

STAFFAN A. FREDRICSSON

Abstract—**Optimization of the transmitting filter in a PAM system using a Viterbi detector of constrained complexity is considered. The receiving filter is considered to be a whitened matched filter. A constraint on detector complexity is obtained by limiting the length of the system impulse response. The results are applied to a channel with coaxial cable characteristics. Comparison with other detectors shows that the Viterbi detector is preferable even when the length of the system impulse response is quite short.**

Manuscript received December 12, 1973. This work was supported by the Swedish Board for Technical Development.

The author is with the Department of Telecommunication Theory, Royal Institute of Technology, Stockholm, Sweden.

I. Introduction

DEMANDS for higher data rates in the transmission of digital information are continuously increasing. In order to mitigate the intersymbol interference effects which inevitably accompany such increased data rates, several different systems have been proposed. In digital PAM systems the use of partial response techniques looks very promising. Recently, Forney [1] and Kobayashi [2] have shown that the so-called Viterbi detector performs maximum-likelihood sequence estimation of the transmitted sequence in partial response systems. The main drawback