and $n \to \infty$ in such a manner that $k/n \to 0$. We have not attempted an investigation of the precise nature of this convergence.

We remark that if we knew that x and θ were jointly Gaussian, we could certainly use knowledge of this fact to develop a more sophisticated estimator than the *NN* estimator. However, the *NN* estimator has the advantage that $R_n^{(k)} \leq 2\sigma_1^2 < \infty$, for all n and k. Thus even a single sample yields finite risk. In contrast, consider a standard linear regression technique utilizing the fact that, in the Gaussian case, $\mu(x)$ is of the form $ax + b$ to form estimates $\hat{a}$ and $\hat{b}$ on the basis of the minimum mean-squared error line fitting the points $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$. The resulting estimator $\hat{\theta} = \hat{a}x + \hat{b}$ has infinite risk for sample sizes of $n = 1, 2, 3$. Thus the nonparametric *NN* estimator is actually better for sufficiently small sample size.

VI. Conclusions

It has been shown that the large sample risk for the *NN* decision rule is no greater than twice the Bayes risk for both the squared-error loss function and the metric loss function. In particular, it has been shown under mild continuity conditions that the conditional risk $r(x)$ of the *NN* estimation rule in the infinite sample case satisfies the inequalities

$$r(x) \leq 2r^*(x), \text{ for the metric loss case, and}$$
$$r(x) = 2r^*(x), \text{ for the squared-error loss case.}$$

Under certain additional moment conditions the unconditional risk R of the *NN* estimate satisfies

$$R \leq 2R^*, \quad \text{for metric loss,}$$
$$R = 2R^*, \quad \text{for squared-error loss, and}$$
$$R = (1 + 1/k)R^*, \quad \text{for squared-error loss with a } k \text{ NN estimate.}$$

These conclusions are complemented by those of earlier work,[4],[5] in which it is shown that $R \leq 2R^*$ for the classification problem with a probability of error loss criterion. Thus the most sophisticated decision rule, based on the entire sample set, may reduce the risk by at most a factor of two. In this sense it may be concluded that at least half the decision information in an infinite set of classified samples is contained in the nearest neighbor.

Acknowledgment

The author wishes to thank B. Efron and P. Hart for their helpful comments during the preparation of this paper.

References

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

[2] ——, "Discriminatory analysis: small sample performance," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 11, August 1952.

[3] M. V. Johns, "An empirical Bayes approach to non-parametric two-way classification," in *Studies in Item Analysis and Prediction*, Herbert Solomon, Ed. Stanford, Calif.: Stanford University Press, 1961.

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

[5] P. E. Hart, "An asymptotic analysis of the nearest-neighbor decision rule," Stanford Electronics Labs., Stanford University, Stanford, Calif., Tech. Rept. 1828-2, May 1966.

On the Mean Accuracy of Statistical Pattern Recognizers

GORDON F. HUGHES, MEMBER, IEEE

Abstract—**The overall mean recognition probability (mean accuracy) of a pattern classifier is calculated and numerically plotted as a function of the pattern measurement complexity n and design data set size m. Utilized is the well-known probabilistic model of a two-class, discrete-measurement pattern environment (no Gaussian or statistical independence assumptions are made). The minimum-error recognition rule (Bayes) is used, with the unknown pattern environment probabilities estimated from the data relative frequencies. In calculating the mean accuracy over all such environments, only three parameters remain in the final equation: n, m, and the prior probability p_c of either of the pattern classes.**

With a fixed design pattern sample, recognition accuracy can first increase as the number of measurements made on a pattern increases, but decay with measurement complexity higher than some optimum value. Graphs of the mean accuracy exhibit both an optimal and a maximum acceptable value of n for fixed m and p_c. A four-place tabulation of the optimum n and maximum mean accuracy values is given for equally likely classes and m ranging from 2 to 1000.

The penalty exacted for the generality of the analysis is the use of the mean accuracy itself as a recognizer optimality criterion. Namely, one necessarily always has some particular recognition problem at hand whose Bayes accuracy will be higher or lower than the mean over all recognition problems having fixed n, m, and p_c.

Manuscript received November 3, 1966; revised July 19, 1967. This work was supported in part by RADC under Contract AF 30(602)-3976.

The author is with Autonetics, a division of North American Rockwell, Inc., Anaheim, Calif. 92803

I. Introduction

SOME consequences of the statistical model of pattern recognition will be presented.[1]–[5] It will be shown that certain useful numerical conclusions can be drawn from rather few assumptions. Basically, the only

assumption made is that some unknown discrete probability structure underlies the pattern environment. This structure defines the particular recognition problem under consideration. A pattern is obtained by making a sample measurement on the environment. Each possible pattern is to be classified into one of a set of classes by a *recognition rule* having maximal probability of correct classification (Fig. 1).

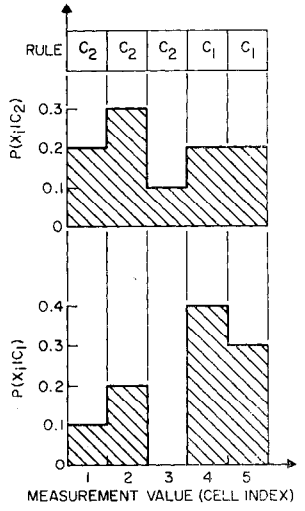


Fig. 1. Known probability example.

It is hoped that this model is sufficiently simple and standard[7],[17] to be readily accepted. The intent of the present analysis is to obtain results which follow from the statistical model itself; e.g., *without* constraining the unknown probabilities to be Gaussian or otherwise parametric. For the same reason, no assumption of statistical independence between pattern measurements will be made. However, only the case of two pattern classes will be treated.

An optimality criterion will be proposed for recognition rules designed from data sets of m sample patterns. The measurement complexity n (total number of discrete values) will also be used as a parameter. It will be shown that it is also desirable to exhibit explicitly the prior probability p_{c1} of class c_1 as the third criterion parameter (where $p_{c2} = 1 - p_{c1}$).

The criterion proposed is the mean correct recognition probability $\bar{P}_{cr}(n, m, p_{c1})$ over all pattern environments. It is obtained by first fixing n, m, and p_{c1} and calculating the accuracy $P_{cr}(n, m, p_{c1})$ of the minimal error-rate Bayes recognition rule for any given environment probability structure.[15] This accuracy is then averaged over all such probability structures, giving the mean accuracy $\bar{P}_{cr}(n, m, p_{c1})$. The resultant equation will be numerically evaluated for practical ranges of n, m, and p_{c1}. Some interesting optimality relationships will thereby be exhibited.

A forewarning should be made of a potential misinterpretation of this criterion. In the literature, the term "mean Bayes accuracy" often refers to averaging only over the possible measurement values of a fixed recognition problem, namely, one having some known environment probability structure.[16] Here, the mean refers to averaging over all recognition problems having fixed measurement complexity n and prior class probability p_{c1}. For example, this includes all subsets of parametric and/or statistically dependent probability structures, as well as all other discrete, normalizable structures. Consequently, the criterion evaluates the *overall* performance of a Bayes recognition rule. Middleton[2] (Section 23.4) presents a useful background discussion on criteria selection.

II. Statistical Model

A statement of the statistical pattern recognition model to be used will first be made. It is fairly standard and contains as few assumptions and constraints as possible. By the term "model" is meant that all the required assumptions are stated, and the applicability of the subsequent analysis to particular recognition problems can be judged on the model itself.

It is assumed that there exist two statistical environments called (caused by) pattern classes c_1 and c_2. Each environment is characterized by a constant but unknown probability distribution over the n discrete values of the pattern measurement variable x. Namely, $P(x_i \mid c_1)$ is the probability of measurement value x_i occurring in environment c_1 for $i = 1, 2, \cdots, n$. These scalars are termed the cell probabilities. Similarily, the unknown distribution $P(x_i \mid c_2)$ characterizes x under the second environment c_2. Vector measurements will be discussed later.

By *unknown*, it is meant that no prior information whatever is given on the $2n$ scalars $P(x_i \mid c_j)$. Any pair of distributions is equally likely a priori. Sample pattern data from the environments is to be measured to estimate the actual $P(x_i \mid c_j)$ existing in any specific recognition problem (the sample relative frequencies will be used).

Thus, the only constraints are that all probabilities be non-negative and that

$$\sum_{i=1}^{n} P(x_i \mid c_1) = \sum_{i=1}^{n} P(x_i \mid c_2) = 1. \tag{1}$$

Whenever a measurement x is made, there is a known prior class probability p_{c1} that class c_1 is in effect and $P(x_i \mid c_1)$ applies. With the complementary probability $p_{c2} = 1 - p_{c1}$, class c_2 and $P(x_i \mid c_2)$ are in effect. However, the particular class in effect for any pattern is not known: only the value x_i produced by the pattern measurement is available.

The pattern recognition problem is to design a recognition rule to predict (recognize) the pattern class most likely in effect for each of the n possible measurement values x_i. Its theoretic solution is known to be a maximal class recognition accuracy Bayes rule.[17] Specifically, this rule is to predict c_1 when x_i occurs if $P(c_1 \mid x_i) > P(c_2 \mid x_i)$;

otherwise predict c_2.[1] This superficially simple rule states to choose the more probable class, given the measurement value x_i which has occurred. The resultant correct recognition probability (accuracy) is then

$$P_{cr}(n, p_{c1}) = \sum_{i=1}^{n} [\max_{j=1,2} P(c_j \mid x_i)]P(x_i) \quad (2)$$

$$= \sum_{i} \max_{j} P(c_j, x_i) \quad (3)$$

$$= \sum_{i} \max_{j} [P(x_i \mid c_j)p_{cj}]. \quad (4)$$

Note that no assumption of measurement statistical independence is made (Appendix I). Also, a vector of r discrete measurements, each having n_k values, $k = 1, 2, \cdots, r$, is clearly equivalent to a single measurement with

$$n = n_1 n_2 \cdots n_r. \quad (5)$$

Appendix I details this equivalence.

Use of discrete measurement values in the model is dictated by practical considerations. Namely, measurements can be made only to a finite precision and consequently only a finite number n of different values can result. Thus n can be termed the *measurement complexity*. Many recognition problems are explicitly digital, such as the visual patterns usually arising in character recognition.[11]

III. Infinite Data Sets

Although the unknown probabilities $P(x_i \mid c_j)$ must actually be statistically estimated from finite pattern sets, the limiting case of known probabilities ($m = \infty$) will be first developed. For example, the two histograms of Fig. 1 give $P(x_i \mid c_1)$ and $P(x_i \mid c_2)$ for a problem having $n = 5$. Given that the classes are equally likely ($p_{c1} = p_{c2} = \frac{1}{2}$), the optimal recognition rule for the five values is as shown, and the recognition accuracy from (3) and (4) is

$$P_{cr}(5, \tfrac{1}{2})$$

$$= P(c_2, x_1) + P(c_2, x_2) + P(c_2, x_3) + P(c_1, x_4) + P(c_1, x_5)$$

$$= 0.65. \quad (6)$$

It is obvious, but still important to note, that this Bayes accuracy of 65 percent is the best possible for the probability structure of Fig. 1. No amount of "improved recognizer design" can increase this figure. Clearly all accuracies will lie between max (p_{c1}, $1 - p_{c1}$) and unity, since the former can be obtained with no measurements whatever (the rule would be always to predict the same class, that having higher prior probability). In fact, it will be shown in (16) that the highest average accuracy ($n = \infty$) is only 75 percent for $p_{c1} = \frac{1}{2}$.

Next, it may be observed that the distributions of Fig. 1 possess no particular continuity, symmetry, or modality over the x_i range. No reasonable parametric forms could be assumed for $P(x_i \mid c_1)$ and $P(x_i \mid c_2)$, such as a pair of Gaussian density functions

$$P(x_i \mid c_j) = N(\mu_j, \sigma_j)$$

$$\doteq (1/\sqrt{2\pi\sigma_j^2}) \exp [-(x_i - \mu_j)^2/2\sigma_j^2].$$

This *nonparametric* aspect of the pattern environment appears to be common to many recognition problems. It is further discussed in Section VII, using some sample histograms from photographic recognition data.

Alternative parametric assumptions have been tried by several workers. These include heuristic fitting of an overlapping sequence of multivariate Gaussian densities to the data.[9] Local smoothness assumptions have been made on the densities, according to a metric which leads to a nearest neighbor recognition rule.[1],[8] However, the continuity constraints which would be imposed on the model by these assumptions are of an entirely different nature than the simple normalizing constraints of (1). Also, the validity of such assumptions is often difficult to verify (see Kendall and Stuart,[10] Section 30.63).

In keeping with the desire for minimal constraints on the model, no such continuity or parametric requirements will be imposed at all. Instead, sample pattern data will be used to estimate individually the discrete cell probabilities by computing relative frequencies.

IV. Evaluation Criterion

A natural criterion to evaluate recognition rule performance is the expected or mean Bayes recognition accuracy over all possible environment probabilities $P(x_i \mid c_j)$. Namely, no prior information on each scalar $P(x_i \mid c_j)$ is to be assumed before the sample pattern data are measured. Any set of $2n$ positive real probability values is equally likely as long as (1) is satisfied.[2]

Clearly, if any such set were made more likely than another, then the criterion would emphasize the accuracy on that particular recognition problem. Instead, the criterion is to weigh equally all recognition problems having given values of p_{c1} and n.

It should be remarked that p_{c1} is explicitly exhibited as a parameter because recognition rule performance should be judged against the minimum accuracy of max (p_{c1}, $1 - p_{c1}$) using *no* measurements. If p_{c1} lies near zero or unity, then any recognizer should have nearly 100 percent accuracy.

Thus the Bayes accuracy of (4) is a statistic, in that it is a function of the random variables

[1] There exist theoretic cases where ties $P(c_1 \mid x_i) = P(c_2 \mid x_i)$ must be more carefully treated, or where randomized decision rules arise.[5] Also, only the direct criterion of maximal recognition accuracy will be used here. No significant change in the analysis results if the misrecognition probabilities are weighted in order to speak of maximizing utility or minimizing risk costs.

[2] This is equivalent to assuming Bayes' postulate (Kendall and Stuart,[10] sec. 8.4). Its use here is fundamentally based on the discrete nature of the distributions.

$$u_i \doteq P(x_i \mid c_1)$$
$$v_i \doteq P(x_i \mid c_2) \qquad i = 1, 2, \cdots, n. \tag{7}$$

To compute its mean or expected value, first note that the u_i and v_i are uniformly distributed due to the "equally likely" assumption of the model:

$$dP(u_1, u_2, \cdots, u_n, v_1, v_2, \cdots, v_n) = N\, du_1\, du_2 \cdots du_{n-1}\, dv_1\, dv_2 \cdots dv_{n-1}. \tag{8}$$

Only $2(n-1)$ differentials appear on the right of (8) because the two normalizing equations (1) fix u_n and v_n in terms of the others. Now, the *boundaries* of the $(n-1)$-order distribution of the u_i alone are given by the intersection of the hypercube $0 \leq u_i \leq 1$, $i = 1, 2, \cdots, n-1$, and the symmetric hyperplane

$$\sum_{i=1}^{n-1} u_i = 1, \tag{9}$$

which is also caused by (1). An identical boundary structure holds for the v_i, so that the normalizing constant N in (8) is obtained from

$$1 = N\left[\int_0^1 du_1 \int_0^{1-u_1} du_2 \int_0^{1-u_1-u_2} du_3 \cdots \cdot \int_0^{1-u_1-u_2-\cdots-u_{n-2}} du_{n-1}\right] \cdot \left[\int_0^1 dv_1 \int_0^{1-v_1} dv_2 \int_0^{1-v_1-v_2} dv_3 \cdots \cdot \int_0^{1-v_1-v_2-\cdots-v_{n-2}} dv_{n-1}\right]. \tag{10}$$

These two iterated integrals may be easily evaluated, giving

$$N = [(n-1)!]^2. \tag{11}$$

Equation (4) is the recognition accuracy given the u_i, v_i and is multiplied by (8) to obtain a joint probability. This is integrated over the u_i, v_i range to get the mean accuracy. By symmetry, each of the n terms of (4) will have an identical expected value, so that n times the expected value of the first may be taken:

$$\bar{P}_{cr}(n, p_{c1}) = n[(n-1)!]^2 \int_0^1 \int_0^1 \max(p_{c1}u_1, p_{c2}v_1)\, du_1\, dv_1 \cdot \left[\int_0^{1-u_1} du_2 \int_0^{1-u_1-u_2} du_3 \cdots \int_0^{1-u_1-u_2-\cdots-u_{n-2}} du_{n-1}\right] \cdot \left[\int_0^{1-v_1} dv_2 \int_0^{1-v_1-v_2} dv_3 \cdots \int_0^{1-v_1-v_2-\cdots-v_{n-2}} dv_{n-1}\right] \tag{12}$$

$$= n(n-1)^2 \int_0^1 \int_0^1 (1-u_1)^{n-2}(1-v_1)^{n-2} \cdot \max(p_{c1}u_1, p_{c2}v_1)\, dv_1\, du_1. \tag{13}$$

By requiring that $p_{c1} \leq p_{c2}$ (without loss of generality), the v_1 integral in (13) may be broken into two ordinary integrals. The first is over the range $0 \leq v_1 \leq p_{c1}u_1/p_{c2}$, and the second requires an integration by parts. The u_1 integral may be then evaluated as a beta function plus a second, somewhat cumbersome integral whose integrand may be expanded by the binomial theorem and integrated term by term, giving

$$\bar{P}_{cr}(n, p_{c1}) = p_{c1} + p_{c2}(n-1)\left(\frac{p_{c1}}{p_{c2}}\right)^n \cdot \sum_{j=0}^{n} \frac{n!}{j!\,(n-j)!\,(2n-j-1)[p_{c1}/(1-2p_{c1})]^j}, \qquad (p_{c1} < p_{c2}). \tag{14}$$

For the common case $p_{c1} = p_{c2} = \frac{1}{2}$, (14) reduces to

$$\bar{P}_{cr}(n, \tfrac{1}{2}) = \frac{3n-2}{4n-2}. \tag{15}$$

Equation (14) is plotted in Fig. 2 for p_{c1} ranging from 0.1 to 0.9 and n from 1 to 1000. Smooth curves have been drawn for clarity, despite the actual discrete measurement nature.

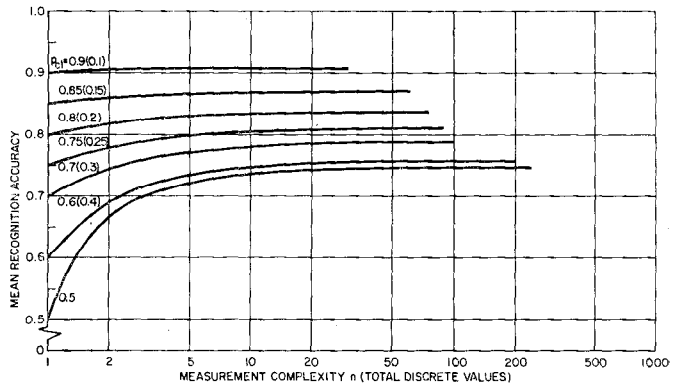


Fig. 2. Infinite data set accuracy.

Note that each recognition accuracy curve begins at $\max(p_{c1}, 1-p_{c1})$ for $n = 1$ (a single measurement value which must always occur and therefore imparts no information). It monotonically increases with the measurement complexity.

Each curve is also near its asymptotic *maximum recognition accuracy* for all $n \gg 2$. This value is *significantly less than* 100 *percent* unless p_{c1} is near zero or unity. In general, the knee of each curve would occur near $n = n_c$, the number of pattern classes defined.

The asymptotic maximum accuracy is obtained by letting $n \to \infty$ in (13):

$$\bar{P}_{cr}(n = \infty, p_{c1}) = p_{c1} + p_{c2}^2 = p_{c2} + p_{c1}^2 = 1 - p_{c1}p_{c2}. \tag{16}$$

V. Finite Data Sets

Equation (14) gives the mean Bayes accuracy in the limiting case where the probability structure is exactly measurable, by means of an infinite set of sample patterns

($m = \infty$). At first glance, it might seem that since only finite data sets exist, the true mean accuracy $\bar{P}_{cr}(n, m, p_{c1})$ could only be approximated. This is because only finite-sample estimates of the true probabilities in (4) can be made. Thus, the statistical model would have to be evaluated by Monte Carlo trials. Experimental sets of testing patterns would be required in order to measure the overall mean accuracy.

It is interesting that this is *not* so. An exact expression for the true mean accuracy can be derived, extending (14) to finite m. Of course this analytic preciseness is obtained at the cost of using a *mean* accuracy criterion. One necessarily always has some *particular* recognition problem at hand with unknown but specific values of $P(x_i \mid c_j)$. It may therefore have an accuracy either higher or lower than the overall mean.

The most direct procedure for employing the m sample patterns to estimate the cell probabilities is to use the *relative frequencies.*[18] If m_1 of the samples are taken under class environment c_1, and measurement value x_1 occurs s times, then the estimate is $P(x_1 \mid c_1) \simeq s/m_1$. Similarly, if r of the $m_2 = m - m_1$ samples taken under c_2 also fall in cell x_1, then $P(x_1 \mid c_2) \simeq r/m_2$.

Such direct use of the relative frequencies is quite natural, as compared to alternate methods using the sample moments.[19] It also offers certain well-known optimality characteristics. For example, the relative frequencies are unbiased and consistent, and also are the maximum-likelihood estimates of the cell probabilities (Appendix I).

To avoid sampling bias, it is necessary that $m_1 = p_{c1}m$ and $m_2 = p_{c2}m$ (truncated to integer values). Consider the first term $i = 1$ in (4). In place of

$$\max [P(x_1 \mid c_1)p_{c1}, P(x_1 \mid c_2)p_{c2}],$$

one has

$$\max \left[\frac{s}{m_1} p_{c1}, \frac{r}{m_2} p_{c2}\right] = \frac{\max (s, r)}{m}. \tag{17}$$

Thus, the *relative frequency recognition rule* for classifying x_1 is to choose c_1 if $s > r$ and c_2 if $s < r$. If $s = r$, choose the class with higher prior probability. Retaining the previous arbitrary assumption $p_{c1} \leq p_{c2}$, choose c_2 for such ties.

In general, the rule for classifying x_i is

$$\text{predict} \begin{cases} c_1 & \text{if} \quad s_i > r_i \\ c_2 & \text{if} \quad s_i \leq r_i \qquad (p_{c1} \leq p_{c2}). \end{cases} \tag{18}$$

Now, the relative frequencies tend to the true probabilities as $m \to \infty$. Also, the Bayes rule accuracy is the maximum attainable for any rule. Consequently, one expects that the accuracy curves for (18) will lie below those of Fig. 2 and rise into coincidence as $m \to \infty$.

Appendix II gives the derivation of the recognition accuracy for finite m; the result is

$$\bar{P}_{cr}(n, m, p_{c1}) = \sum_{r=0}^{m_2} \sum_{s=0}^{m_1} \left[\frac{(m_2 - r + 1)(m_2 - r + 2) \cdots (m_2 - r + n - 2)}{(m_2 + 1)(m_2 + 2) \cdots (m_2 + n - 1)}\right] \cdot \left[\frac{(m_1 - s + 1)(m_1 - s + 2) \cdots (m_1 - s + n - 2)}{(m_1 + 1)(m_1 + 2) \cdots (m_1 + n - 1)}\right] g(r, s)$$

where

$$g(r, s) = n(n-1)^2 \cdot \begin{cases} p_{c1} \dfrac{s+1}{m_1 + n} & \text{for} \quad s > r \\ p_{c2} \dfrac{r+1}{m_2 + n} & \text{for} \quad s \leq r. \end{cases} \tag{19}$$

Equation (19) is best numerically evaluated by digital computer (compare Appendix II), and is plotted in Fig. 3 for the common case of $p_{c1} = p_{c2} = \frac{1}{2}$, n ranging from 1 to 1000, and m ranging from 2 to 1000. The limiting $m = \infty$ curve from Fig. 2 has been appended. Smooth curves have again been drawn for clarity.

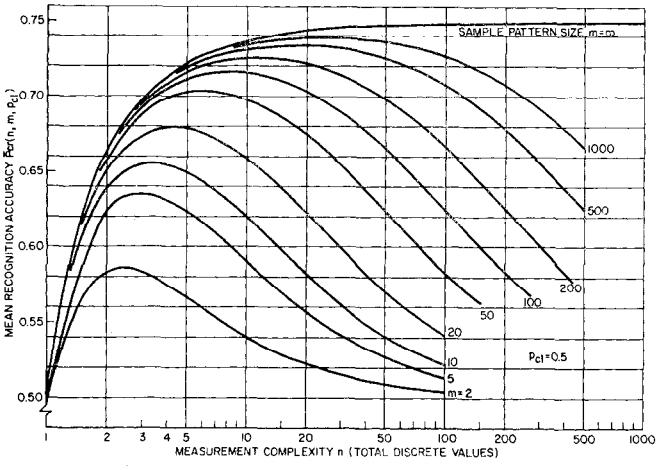


Fig. 3. Finite data set accuracy ($p_{c1} = \frac{1}{2}$).

The most interesting aspect of Fig. 3 is the existence of an *optimal measurement complexity* n_{opt} which maximizes the mean accuracy for any given $m < \infty$. Examples are $n_{opt} = 17$ for $m = 500$ and $n_{opt} = 23$ for $m = 1000$ (an interpretation of the relative magnitude of n_{opt} will be made in Section VIII). This behavior is quite reasonable. With m fixed, the accuracy first begins to rise with n as in Fig. 2. It ultimately must fall back as $n/m \to \infty$ because the precision of the probability estimates monotonically degrades. It is shown in Appendix I that this precision degrades as $\sigma(s/m_1)/p(x|c_1) \alpha \sqrt{n/m}$. However, the effect is easy to see when $n \gg m$ because most of the $2n$ cells will contain no samples at all. This then gives zero relative frequencies for these cells, irrespective of the actual cell probabilities. Rule (18) then consists mostly of $r_i = s_i = 0$ ties, which are all resolved by choosing c_2.

Consequently, as n varies from one to infinity, one expects the accuracy to rise to a maximum, fall back, and asymptotically approach max (p_{c1}, p_{c2}). Table I presents the optimal values for equally probable classes.

TABLE I

OPTIMAL MEASUREMENT COMPLEXITY

m	n_{opt}	$\bar{P}_{cr}(m, n_{opt}, \frac{1}{2})$
2	2	0.5833
5	3	0.6350
10	3	0.6548
20	4	0.6791
50	6	0.7031
100	8	0.7161
200	11	0.7257
500	17	0.7345
1000	23	0.7390
∞	∞	0.7500

VI. Unequal Class Probabilities

Equation (19) may also be used to illuminate what appears to be a critical aspect of Bayes recognizers designed from finite sample sets. Namely, if one pattern class is appreciably more likely to occur than another, then there exists a maximum acceptable measurement complexity n_{max}. Its value increases with m. Specifically, for $n > n_{max}(m)$, the recognizer is inferior to simple random guessing (i.e., always predicting the more likely class c_2 irrespective of the measurement value occurring). Note that the actual value of p_{c2} need not be known here, only that it exceeds p_{c1}.

Thus, the general statement is that $\bar{P}_{cr}(n, m, p_{c1}) <$ max (p_{c1}, p_{c2}) for all n exceeding some n_{max}, if p_{c1} sufficiently differs from p_{c2}.

Rather than present an algebraic proof of this statement, it seems more useful to exhibit the effect directly using (19). Fig. 4 presents the accuracy curves for the case of $p_{c1} = 0.2$. The maximum values $n_{max}(m)$ occur where the curves fall below the level $\bar{P}_{cr} = 0.8$. Note that $n_{max}(m) \sim m/2$. Graphically, the curves are similar to Fig. 3 except that the asymptote 0.80 is approached from below. This random-guessing asymptote at $n = \infty$ is due to the ties in (18).[3] An interpretation of the relative magnitude of n_{max} will be made in Section VIII.

It is not difficult to state the reason for this behavior: *If insufficient sample data are available to estimate the pattern probabilities accurately, then a Bayes recognizer is not necessarily optimal.* Stated this simply, of course, the statement is obvious. None the less, the problem is often ignored. In addition, the statement remains qualitatively valid if the probabilities are assumed to be parametric (e.g., Gaussian). The problem is most severe when one pattern class dominates (p_{c2} near unity), and vanishes when both classes are equally likely.

[3] The minimum of each curve thus gives the worst case of a recognizer which is too complex for its design pattern set. In practice, a recognizer inferior to simple random guessing should be rejected. The asymptote line $P_{cr} = 0.80$ would thus apply for all $n > n_{max}$.

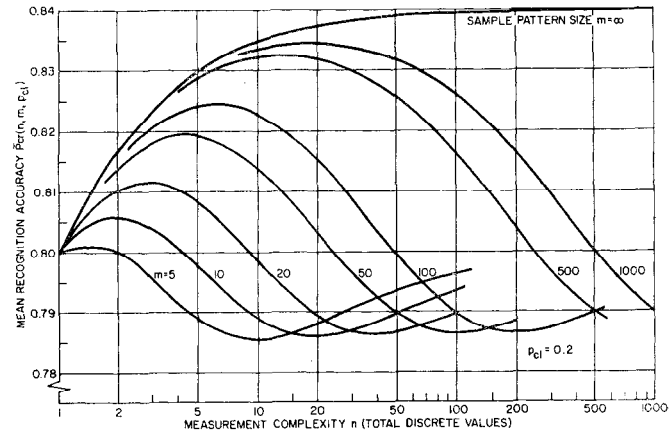


Fig. 4. Finite data set accuracy ($p_{c1} = \frac{1}{5}$).

VII. Experimental

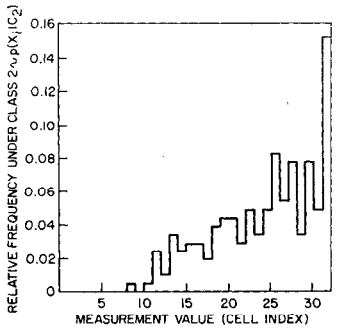
Fig. 5 presents relative frequency (histogram) data gathered from an aerial photograph processor. It was designed to distinguish industrial areas (c_1) from residential urban areas (c_2). The pattern measurement made is the edge density. An edge is defined to exist at any point on the photograph where the two-dimensional intensity-gradient magnitude $|\nabla I|$ exceeds a threshold value. Note that an edge so defined is not necessarily part of a continuous boundary.

Essentially this measurement is able to distinguish between the two classes because homes tend to be small and closely packed, compared to industrial areas. The latter typically contain fewer and larger structures, as well as low-detail areas such as storage yards or parking lots. Consequently, the residential histogram of Fig. 5(a) is concentrated towards the higher edge density values. It is fairly disjoint from the industrial histogram.

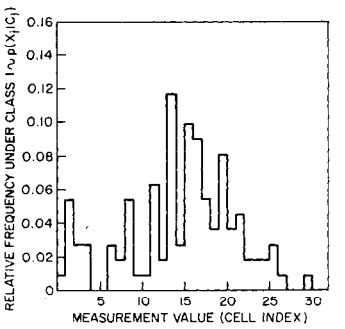
Take these histograms as estimates of the cell probabilities and assume $p_{c1} = p_{c2} = \frac{1}{2}$. The Bayes accuracy estimate from (4) is $P_{cr} \cong 0.787$, calculated as in (6). A total of $m = 315$ sample patterns was used in the histograms, and there are $n = 32$ measurement values. From Fig. 3, the mean accuracy is then $\bar{P}_{cr}(32, 315, 0.5) = 0.723$, somewhat below the particular value of 0.787 obtained. Although the choice of 32 measurement values is higher than the optimal value of 13 (Table I), the mean accuracy is not appreciably less than its maximum of 0.729.

As a final experimental topic, one can show that this photographic recognition problem is indeed discontinuous and non-Gaussian. More broadly, it cannot be fitted by reasonable parametric probability functions. This conclusion rests on the large histogram discontinuities shown in Fig. 5; for example, the 210 percent increase from cell 31 to cell 32 under c_2. Or note the 77 percent decrease from cell 14 to cell 15 under c_1.

However, it must be shown that these discontinuities reflect the actual underlying probabilities. They could conceivably be caused merely by random variations due to the finite sample size of 315.



(a)



(b)

Fig. 5. Aerial photograph processor histograms.

A rough argument usffices to dispose of this possibility. It is well known that the relative frequency r_i/m_1 for each cell of c_1 (say) is a binomially distributed random variable (Appendix I). Its expected value is equal to the true cell probability u_i, and its sampling variance is

$$\begin{aligned}\sigma_i^2 &= u_i(1 - u_i)/m_1 \\ &= r_i(m_1 - r_i)/m_1^3 + O(m_1^{-5/2}).\end{aligned} \tag{20}$$

It is unlikely[4] that the histogram values r_i/m_1 will deviate from the true cell probabilities by more than σ_i. In Fig. 5(a), $m_1 \cong 315/2$ and cell 32 is thus unlikely to have true probability less than $0.152 - \sigma_{32} \cong 0.123$. Cell 31 is unlikely to be more probable than $0.049 + \sigma_{31} \cong 0.066$. Therefore, it is highly likely that a significant discontinuity exists between p_{31} and p_{32} themselves. Roughly, "highly likely" means having probability about $1 - (0.159)^2 = 0.975$.

A more precise binomial argument gives Pr $[p_{32} < 0.123] = 0.118$ and Pr $[p_{31} > 0.066] = 0.271$. The discontinuity likelihood is thus nearly $1 - (0.118)(0.271) = 0.968$, rather than 0.975.

VIII. Conclusions

Certain results have been drawn from the statistical model of pattern recognition by using mean value arguments. Most interesting is the existence and size of an optimal measurement complexity, n_{opt}, as well as a maximum acceptable complexity.

If all pattern probabilities were known in advance, the recognition accuracy would be expected to be nearly maximal if $n \gg n_c = 2$ (n_c being the number of pattern classes). More than (say) 20 possible measurement values gives little additional aid in making a simple decision between two dichotomous classes.

The recognition accuracy curves confirm this intuitive feeling. In Fig. 2, the infinite data set curves have substantially reached their asymptotes at $n = 20$. In Fig. 3, n_{opt} ranges from 2 to 23 for $2 \leq m \leq 1000$ sample patterns.

On the other hand, one can easily envision, with Marill and Green,[7] a vector pattern having 10 component measurements of 10 possible values each. From (5), this gives $n = 10^{10}$, or some twenty billion cell probabilities to estimate, merely to classify patterns into two classes. How is one to reconcile this with (say) the optimal value of $n = 23$ if only 1000 sample patterns are available?

The answer appears to be that only 23 values are statistically significant. A (singular) mapping should be sought to transform the 10^{10} values into 23 before using the sample patterns to compute the recognition rule. In other words, the problem arises because the original measurement structure is incorrectly defined.

Sometimes this mapping can be found by reconsidering the physical origin of the recognition problem at hand. However, statistical assistance in *measurement selection* is often required. Shannons' information measure[2],[12] or Kullbacks' divergence measure[13] might be evaluated for each of the ten pattern vector measurements, and the lowest scoring five discarded.[13],[14] This would leave $n = 10^5$. More directly, score on the estimated accuracy is derived by inserting (17) into (4).

Next, individual *measurement reduction* may be performed on the five remaining measurements. Suppose the ten values of each are reduced to two by forming individual recognition rules. Since the five rules will usually disagree, a final *measurement combination* of the 2^5 values must be made. This value of 32 is not far above the optimum of 23, so that a combining recognition rule can be computed by (18) for a final class prediction.

It should be emphasized that these examples of mapping by measurement selection, reduction, and combination are not proposed as developed techniques. Rather, they are illustrative of a framework for further investigation. Also of interest would be an extension of the present analysis to $n_c > 2$ classes. This is of no conceptual difficulty, yet a general equation giving $\bar{P}_{cr}(n, m, n_c, p_{c1})$ for all n_c has not been found.

Appendix I

Vector Measurements, Statistical Independence, and Sampling Formulas

A discrete vector-pattern measurement $y = (y^r, y^{r-1}, \cdots, y^2, y^1)$ may be nonsingularly mapped to its equivalent

[4] Using the Gaussian approximation to the binomial distribution,[6],[11] the probability of a sampling deviation more than σ_i in *one* direction is about 0.159. It is 0.023 for more than a $2\sigma_i$ deviation.

scalar measurement x by the x_i cell index equation

$$i = j_1 + n_1(j_2 - 1) + n_1n_2(j_3 - 1) + n_1n_2n_3(j_4 - 1) + \cdots + n_1n_2 \cdots n_{r-1}(j_r - 1). \quad (21)$$

Here, j_k is the cell index of component y^k, $1 \leq j_k \leq n_k$, for $k = 1, 2, \cdots, r$. This mapping is quite useful for computer measurement handling, and is inverted by subtracting unity from i and then successively dividing by $n_1, n_2, \cdots, n_r$; extracting the remainders as $j_1 - 1$, $j_2 - 1, \cdots, j_r - 1$. Indeed, (21) expresses the representation of $i - 1$ in a compound-radix number system having digit radii $n_1, n_2, \cdots, n_r$. Equation (5) follows directly.

To illustrate the probability constraints caused by assuming statistical independence[6] among the measurements y^k, let $r = n_1 = n_2 = 2$. Assume that the component measurements y^2, y^1 are statistically independent under *each* class environment; viz., $P(y^2, y^1 \mid c_i) = P(y^2 \mid c_i) P(y^1 \mid c_i)$ for $i = 1, 2$ and the two values each of y^2 and y^1. Using mapping (21), this leads to the two probability constraints

$$P(x_1 \mid c_i)P(x_4 \mid c_i) = P(x_2 \mid c_i)P(x_3 \mid c_i), \quad (i = 1, 2). \quad (22)$$

This type of constraint would have to be imposed in addition to (1).

Next, the sampling distributions used in the text and Appendix II will be briefly developed. (This is largely a collection of scattered standard results.) Take environment c_1 of the model and let $u_i = P(x_i \mid c_1), i = 1, 2, \cdots, n$. Of m independently sampled patterns, the probability of obtaining a combination with r_i samples in cell i ($i = 1, 2, \cdots, n$) is clearly multinomial[10],[11]:

$$P(r_1, r_2, \cdots, r_n \mid m, u_1, u_2, \cdots, u_n) = \frac{m!}{r_1!\, r_2! \cdots r_n!} u_1^{r_1} u_2^{r_2} \cdots u_n^{r_n} \quad (23)$$

where

$$\sum_{i=1}^{n} r_i = m. \quad (24)$$

For an individual relative frequency r_i/m, the distribution is binomial, since there is probability u_i of falling in cell i and $1 - u_i$ of falling elsewhere:

$$P(r_i \mid m, u_i) = \frac{m!}{r_i!\,(m - r_i)!} u_i^{r_i}(1 - u_i)^{m-r_i}. \quad (25)$$

The relative frequency can thereby be easily shown[11] to have expected value

$$\mu(r_i/m) = u_i \quad (26)$$

and variance

$$\sigma^2(r_i/m) = u_i(1 - u_i)/m. \quad (27)$$

Since the expected value is equal to u_i, the relative frequency is an *unbiased* estimator of u_i. Since the variance vanishes at $m = \infty$, it is also *consistent*. The maximum-likelihood estimator $\hat{u}_i$ of u_i is obtained from the likelihood equation[11]:

$$\frac{\partial[\log P(r_i \mid m, \hat{u}_i)]}{\partial \hat{u}_i} = 0. \quad (28)$$

Using (25), this shows that the *relative frequency is the maximum-likelihood estimator* $\hat{u}_i = r_i/m$.

For a measure of the precision of the relative frequency estimator, one may take σ/μ from (26) and (27):

$$\frac{\sigma(r_i/m)}{\mu(r_i/m)} \sim \sqrt{n/m}. \quad (29)$$

Here, $u_i \sim 1/n$ has been used to get an order of magnitude.

Appendix II

Derivation of Accuracy with Finite Data Set

The derivation of (19) begins analogously to that of (14), except (17) is used instead of the max function in (4). As before, the contribution of measurement value x_1 is equal to any other by symmetry, giving

$$\bar{P}_{cr}(n, m, p_{c1}) = nE\{u_1 p_{c1} P[s_1 > r_1 \mid u_1, v_1] + v_1 p_{c2} P[s_1 \leq r_1 \mid u_1, v_1]\}, \quad (p_{c1} \leq p_{c2}) \quad (30)$$

where $u_1 = P(x_1 \mid c_1)$, $v_1 = P(x_1 \mid c_2)$. Now s_1 and r_1 arise independently from m_1 samples of environment c_1 and m_2 of c_2 (m_1 is the integral part of $p_{c1}m$ and $m_2 = m - m_1$). Consequently, the joint probability of s_1 and r_1 is the product of two binomial distributions of the form (25):

$$P[s_1, r_1 \mid u_1, v_1] = \frac{m_1!\, m_2!}{s_1!\,(m_1 - s_1)!\, r_1!\,(m_2 - r_1)!} \cdot u_1^{s_1}(1 - u_1)^{m_1 - s_1} v_1^{r_1}(1 - v_1)^{m_2 - r_1}. \quad (31)$$

The cumulative probability of $s_1 > r_1$ is obtained by simply summing (31) over all r_1, s_1 values obeying the inequality. Interchanging this linear operation with the expectation operator in (30) (representing a $2(n - 1)$-fold integration over the u_i, v_i ranges as in (10), the expected value of a single term of the $s_1 > r_1$ variety may be seen to be

$$n(n - 1)^2 p_{c1}(s_1 + 1) \cdot \left[\frac{(m_2 - r_1 + 1)(m_2 - r_1 + 2) \cdots (m_2 - r_1 + n - 2)}{(m_2 + 1)(m_2 + 2) \cdots (m_2 + n - 1)}\right] \cdot \left[\frac{(m_1 - s_1 + 1)(m_1 - s_1 + 2) \cdots (m_1 - s_1 + n - 2)}{(m_1 + 1)(m_1 + 2) \cdots (m_1 + n)}\right]. \quad (32)$$

A term of the $s_1 \leq r_1$ variety is obtained from this by interchanging c_2 and c_1, s_1 and r_1, m_1 and m_2, and the final form of (19) then directly follows.

Numerical evaluation of (19) requires either a three-parameter tabulation or a digital computer. For $n =$

$m = 1000$ in Fig. 3, there are about 10^9 multiplications in the four continued products which range up to magnitude $1500!/500! \cong 10^{2980}$. Consequently, a numerical tabulation is the more reasonable approach (four-place tables for $p_{c1} = \frac{1}{5}$ and $\frac{1}{2}$ are available from the author).

References

[1] C. K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electronic Computers*, vol. EC-6, pp. 247–254, December 1957.

[2] D. Middleton, *Statistical Communication Theory*. New York: McGraw-Hill, 1960.

[3] R. G. Miller, "Statistical prediction by discriminant analysis," *Meteorological Monographs*, vol. 4, no. 25, October 1962.

[4] M. Minsky, "A selected descriptor-indexed bibliography to the literature on artificial intelligence," *IRE Trans. Human Factors in Electronics*, vol. HFE-2, pp. 39–56, March 1961.

[5] A. Wald, *Statistical Decision Functions*. New York: Wiley, 1950.

[6] W. B. Davenport and W. L. Root, *The Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958.

[7] T. Marill and D. M. Green, "Statistical recognition functions and the design of pattern recognizers," *IRE Trans. Electronic Computers*, vol. EC-9, pp. 472–477, December 1960.

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

[9] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition*. New York: Macmillan, 1962.

[10] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, 2 vols. New York: Hafner, 1958.

[11] H. Cramér, *Mathematical Methods of Statistics*. Princeton, N. J.: Princeton University Press, 1946.

[12] P. M. Lewis, "The characteristic selection problem in recognition systems," *IRE Trans. Information Theory*, vol. IT-8, pp. 171–178, February 1962.

[13] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. Information Theory*, vol. IT-9, pp. 11–17, January 1963.

[14] J. A. Lebo and G. F. Hughes, "Pattern recognition preprocessing by similarity functionals," *1966 Proc. Nat'l Electronics Conf.*, vol. 22.

[15] Davenport and Root,[6] sec. 14-2.

[16] Middleton,[2] sec. 19.1-1.

[17] Davenport and Root,[6] ch. 14.

[18] Davenport and Root,[8] secs. 5-5.

[19] Cramér,[11] sec. 17.6.

Studies of Sequential Detection Systems with Uncertainty Feedback

JEREMIAH F. HAYES, MEMBER, IEEE

Abstract—The problem studied is the design of signals in a binary sequential detection system which has a feedback channel available to it. Until a decision is made on a particular transmission, the transmitter is informed via the feedback channel of the state of uncertainty at the receiver. This uncertainty feedback modulates the transmitted signal so that, for a prescribed probability of error, the average time of transmission of a binary digit is minimized subject to an average power constraint and a probabilistic peak power constraint.

It is assumed that the forward and the feedback channels are coherent and are each disturbed independently by white Gaussian noise. Previous work has been predicated on noise-free feedback.

Signals are derived, drawn from a certain class, which are optimum when the ratio of allowable peak-to-average power is either large or close to one. The performance of these signals is analyzed. It is shown that for a high signal-to-noise ratio in the feedback channel and a high allowable peak-to-average power ratio one can transmit at approximately channel capacity with low probability of error. However, an infinite feedback channel signal-to-noise ratio and an infinite allowable peak-to-average power ratio are required to transmit at channel capacity with zero probability of error.

Results on signal design in a system whose configuration is slightly different from that considered in the main body of the paper are summarized, as well as the results of a study of the effect of delay in the forward and feedback channels.

Manuscript received September 9, 1966; revised May 22, 1967. This work was supported by the Air Force Office of Scientific Research under Grant AF-AFOSR-230-64.

The author was with the Electronics Research Laboratory, University of California, Berkeley, Calif. He is now with the Dept. of Elec. Engrg., Purdue University, Lafayette, Ind. 47907

Introduction

RECENT WORK of Turin[1] and Horstein[2] has been devoted to the problem of signal design in a binary sequential detection system which has a noiseless feedback link available to it. We consider here the case where the noiseless feedback assumption is no longer justified. Our objective is the same as the former case: to minimize the average time of transmission of a binary digit for a given probability of error, subject to peak and average power constraints on the transmitted signal.[3]

Consider the system depicted in Fig. 1. The forward and the feedback channels are coherent and are each disturbed independently by white Gaussian noise. The double-ended power spectral densities of the noise voltages in the forward and the feedback channels are $N_{01}/2$ and $N_{02}/2$, respectively.

As indicated in Fig. 1, the equally probable outputs of the message source are encoded into the signals $s_{\pm}(w(t), t)$, respectively. At time $t = 0$ the transmitter begins sending the signal corresponding to the current output of the