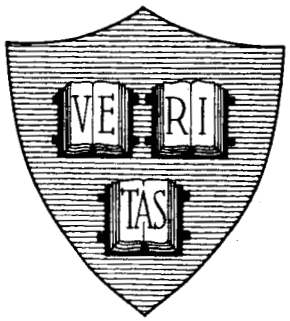
Office of Naval Research

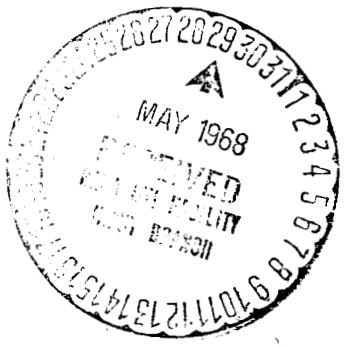
Contract N00014-67-A-0298-0006

NR-372-012

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Grant NGR 22-007-068

ON PATTERN CLASSIFICATION ALGORITHMS--
INTRODUCTION AND SURVEY





By

Yu-Chi Ho and Ashok K. Agrawala

March 1968

Technical Report No. 557

Division of Engineering and Applied Physics
Harvard University • Cambridge, Massachusetts

Office of Naval Research

Contract N00014-67-A-0298-0006

NR - 372 - 012

National Aeronautics and Space Administration

Grant NGR 22-007-068

ON PATTERN CLASSIFICATION ALGORITHMS – INTRODUCTION AND SURVEY

By

Yu-Chi Ho and Ashok K. Agrawala

Technical Report No. 557

March 1968

The research reported in this document was made possible through support extended to the Division of Engineering and Applied Physics, Harvard University by the U. S. Army Research Office, the U. S. Air Force Office of Scientific Research and the U. S. Office of Naval Research under the Joint Services Electronics Program by Contracts N00014-67-A-0298-0006, 0005, and 0008 and by the National Aeronautics and Space Administration under Grant NGR 22-007-068.

Division of Engineering and Applied Physics

Harvard University Cambridge, Massachusetts

ON PATTERN CLASSIFICATION ALGORITHMS — INTRODUCTION AND SURVEY*

By

Yu-Chi Ho and Ashok K. Agrawala

Division of Engineering and Applied Physics

Harvard University Cambridge, Massachusetts

ABSTRACT

This paper attempts to lay bare the underlying ideas used in various pattern classification algorithms reported in the literature. It is shown that these algorithms can be classified according to the type of input information required and that the techniques of estimation, decision, and optimization theory can be used to effectively derive known as well as new results.

---

* The research reported in this work represents an expanded version of a talk of the same title given by the first author at the Tenth Anniversary Seminar of the Statistical Department, Harvard University, April 1967.

I. Introduction

Pattern classification or recognition covers an extremely broad spectrum of problems. Most of us are only concerned with one or two of these at any given time. For example, there is the engineering aspect of the pattern classification problem which is mainly concerned with the implementation and design of actual recognition devices. At the other extreme, there is the artificial intelligence aspect of the problem which is concerned with the philosophical question of learning and intelligence. It is both stimulating and controversial [32, 19]. Similarly, the study of recognition mechanisms in biological systems is another accepted field of study [20]. In this paper, we shall not touch on any of the above mentioned areas. Our survey will be concentrated on what might be called the analytical aspects of the pattern classification problem. By this we mean that the problem is viewed as one of making decisions under uncertainty and the mathematical techniques of decision, estimation, and optimization theory are brought to bear on the problem.

As it is usually understood, there are two fundamental problems associated with this aspect of pattern classification.

(i) Characterization Problem. Given a pattern, signal or waveform, before any decision can be made concerning the pattern, it is often convenient as well as necessary to convert the pattern, signal, or waveform into a set of features or attributes which characterize the pattern under consideration. These features are usually denoted by the real variables $x_1, \ldots, x_m$ and the vector x is called the pattern vector. If we represent the original scanned pattern or sampled

waveform as a vector z, then the characterization or feature selection problem can be simply but vaguely stated as finding a map from z to x, i.e.,

$$x = \phi(z) \tag{I-1}$$

such that x "adequately characterizes" the original z for purposes of classification but the dimension of x is much smaller than that of z.

(ii) Abstraction and Generalization Problem. Once a set of features has been selected, and certain data concerning the patterns and their features are given, the next problem is the determination of a decision function of these features based on the data given such that

$$f(x) = \begin{cases} \geq 0 & x \in \text{class } H^1 \\ < 0 & x \in \text{class } H^0 \end{cases} \tag{I-2}^*$$

The problem of abstracting the necessary information from the given data to produce the decision function f(x)** is called abstraction. Often times, it is convenient but not absolutely necessary to process the given data sequentially or iteratively in order to determine f(x). This iterative procedure for calculating f(x) is known as 'training procedure', 'adaptation', or 'learning.'† Once a decision function f(x) has been found, the generalization problem attempts to assess the goodness of the f(x) through the determination of various error probabilities. Fundamental to this assessment is the knowledge (given

---

* In the main, we shall restrict ourselves to two-class problems. In section IX, we shall discuss the extension to multi-class problems.

** Sometimes f(x) is also referred to as a decision surface in n dimensional x space.

† In the context of this paper, they simply represent entrenched terminology. No philosophical or metamathematical meaning should be attached to these words.

or calculated) of the quantity $P(H^1/x) = 1 - P(H^0/x)$. In fact, the generalization problem can be viewed simply as that of the determination of $P(H^1/x)$.

The distinction between problems (i) and (ii), of course, is not always clear cut; nor can their solutions always be separately considered. For example, how well the characterization problem is solved clearly affects the success of an abstraction algorithm and the ability of the resultant decision function to generalize. In fact, it is generally recognized that (i) is really the principal problem in pattern recognition.

The present paper is devoted to a survey of the various algorithms for the solution of the abstraction problem of pattern classification only. Without minimizing their importance, the characterization and generalization problems will be discussed only to the extent that they are related to the abstraction problem.

II. Types of Input Data

The various abstraction algorithms to be discussed require different types of input data. In this section we shall list these and establish a common notation to be used in the rest of the paper.

There are two pattern classes, H^1 and H^0. The probability of occurence of patterns from the i^{th} class is denoted by $P(H^i)$. If this probability is not explicitly given, then we shall assume it to be equal to $1/2$, i.e., both classes occur equally often. The pattern vectors will be denoted by x with the understanding that the components x_i are features determined as a result of the solution of the characterization problem. Four types of data concerning x will be considered.

(i) Functional form of the condition density $p(x/H^i, \theta)$.

By this we mean that the form of the conditional density functions of x for both classes is given to within the specification of a set of parameters θ. For example, we may be given that the pattern vectors from both classes are gaussianly distributed with unknown mean and covariances.

(ii) Parameters of $p(x/H^i, \theta)$.

By this it is meant that the values of the parameters θ in (i) are also known.

(iii) Sample patterns with known classification.

As part of the given data for the abstraction problem, one is often supplied with a set of training sample patterns of known classification. We denote the two sets

$$\{x^1(1), x^1(2), \ldots, x^1(n_1)\} \triangleq \chi^1(n_1)$$

$$\{x^o(1), x^o(2), \ldots, x^o(n_o)\} \triangleq \chi^o(n_o) \quad .$$

In this case we have two sets of n_1 and n_o samples for classes H^1 and H^o, respectively. For notational compactness, the two sets are often joined to make a matrix, each row of which is a sample pattern from one of the two classes as shown below:

$$A = \begin{bmatrix} 1 & x^1(1)^T \\ 1 & x^1(2)^T \\ \vdots & \\ -1 & -x^o(1)^T \\ -1 & -x^o(n)^T \end{bmatrix} \qquad (n_1 + n_o) \times (1 + m) \text{ matrix}$$

The first column of ones and minus ones is used to indicate the known classification of the patterns.

(iv) Samples of unknown classification.

In so-called problems of training without a teacher, sample patterns of unknown classification are given. In this case they are simply indicated as

$$\chi(n) \triangleq \{x(1), x(2), \ldots, x(n)\}$$

In connection with (iii) and (iv) it is always assumed that the samples are independently chosen. The order of appearance of these patterns is of no significance.

Depending on the combinations of (i)-(iv) that are supplied, different abstraction algorithms result. The following sections will classify and discuss the various algorithms on the basis of these input data and the natural mathematical techniques used in each case.

III. Case A - Data Type (i) And (ii) Are Given

When the conditional density functions $p(x/H^i, \theta)$ including the values of θ are given, the problem reduces to that of simple hypothesis testing in statistics. The basic quantity of interest here is the <u>likelihood ratio</u> defined as

$$L(x) = \frac{p(x/H^i)}{p(x/H^o)} \qquad \text{(III-1)}$$

A decision function formed by comparing L(x) against a threshold value η, i.e.

$$f(x) = L(x) - \eta \qquad \text{(III-2)}$$

is known to be optimal for a variety of criteria depending on the specific value of η. For example, (Selin Ch. 2, 1965 [43]).

(i) Neyman-Pearson Criterion.

Let $X^1 \triangleq \{x \mid f(x) \geq o\}$, and $X^o \triangleq \{x \mid f(x) < o\}$ and

$$\alpha \triangleq \int_{X^1} p(x/H^o)dx = \text{error probability of type 1} \qquad \text{(III-3)}$$

$$\beta \triangleq \int_{X^o} p(x/H^1)dx = \text{error probability of type 2} \qquad \text{(III-4)}$$

If we selected the value of η in (III-2) to yield a fixed value of α, then the decision function f(x) has the property that it minimizes the value of β as compared to any other f(x) yielding the same or smaller α .

(ii) Bayes Criterion.

If the prior probabilities of occurence of the two classes, $P(H^i)$, as well as the cost of wrong decision C_1 and C_2 for the two error types, are given, then selecting

$$\eta = \frac{P(H^o)C_1}{P(H^1)C_2} \qquad \text{(III-5)}$$

will minimize the average risk of making wrong decisions.

(iii) Minimax Criterion.

If the prior probabilities, $P(H^i)$, are unknown, then we may wish to choose the value of η so as to minimize the average risk against the worst value of $P(H^i)$. This is given implicitly by

$$C_1\alpha = C_2\beta \qquad \text{(III-6)}$$

Special Case of the Gaussian $p(x/H^i, \theta)$.

In the case when $p(x/H^i, \theta)$ are gaussian, the likelihood ratio can be explicitly written in terms of the means, μ_i, and covariances

Σ_i, $i = o, 1$. Since the logarithm function is monotone, it is also customary to write $f(x) = \ln L(x) - \ln \eta$ and we have

$$f(x) = \frac{1}{2}[(x - \mu_o)^T \Sigma_o^{-1}(x - \mu_o) - (x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)]$$

$$+ \frac{1}{2} \ln \frac{|\Sigma_o|}{|\Sigma_1|} - \ln \eta \qquad \text{(III-7)}$$

i.e. the optimal decision functions are quadratic. If we furthermore assume that $\Sigma_1 = \Sigma_o = \Sigma$, then Eq. (III-7) simplifies to

$$f(x) = x^T \Sigma^{-1}(\mu_1 - \mu_o) + \text{constant term}$$

$$\triangleq a^T x + a_o \qquad \text{(III-8)}$$

a linear decision function.* In communication terminology, we let components of x represent the successive samples of an input waveform which may be a known signal plus noise or noise only. The components of "a" are then the impulse response of a linear discrete "matched filter" whose output at a given time is the value of the decision function f(x). This is the solution to the problem of detecting the presence of a known signal in gaussian noise.

A linear decision function of the type of Eq. (III-8) also arises naturally in other pattern classification approaches to be described later. Their ease of implementation is a major factor of their popularity. In fact, one is often led to consider only the determination of the best linear decision function based on the given input data. For the gaussian case discussed here, this question has been resolved by Anderson and Bahadur (1962).[6]

---

* "a" here is not to be confused with the a of Eq. (III-3).

Other Optimal Quadratic f(x)

A quadratic f(x) of the type of Eq. (III-7) is actually optimal* for the more general type of distributions than gaussian. Some of these generalizations have been studied by Cooper [14, 13]. Consider the case where $p(x/H^i)$ is given by

$$p(x/H^i) = A_i |\Sigma_i|^{-\frac{1}{2}} h[(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)] \qquad \text{(III-9)}$$

where h is a function integrable in n-space and monotone, i.e. $h(\alpha)$ decreases monotonically for increasing $\alpha, o \leq \alpha < \infty$, and A_i is a constant adjusted to insure $\int p(x/H^i)dx = 1$. It can be shown that μ_i and Σ_i are the mean and covariance matrices respectively of $p(x/H^i)$ and that the class encompasses a wide range of distributions including the normal, Laplace, and rectangular distributions. In the special case when the determinants $|\Sigma_1| = |\Sigma_0|$, then the optimal f(x) is the optimal separating surface for spherical normal, Pearson II and VII types of distributions [13].

Sequential Decision Procedures

In many classification problems, the features or attributes of a sample pattern, x_i, are received sequentially in a natural way, e.g. the x_i's are the sampled value of a waveform in time. In other cases, it may be advantageous to arrange to examine the features in decreasing order of significance with the hope that a classification can be reliably made without having to go through all the features of a pattern most of the time. In either case, one is led to the consideration of sequential decision functions.

---

* In the sense of Bayes criterion with equal cost of misclassification.

The main tool used here is the sequential probability ratio test (SPRT) developed by Wald [49]. This is a natural extension of Eq. (III-1). Let

$$L_j(x) \triangleq L(x_1, \ldots, x_j) = \frac{p(x_1, \ldots, x_j/H^1)}{p(x_1, \ldots, x_j/H^0)} \qquad \text{(III-10)}$$

Instead of a binary choice of decisions after j features, we use the following analog of Eq. (III-2).

$$\begin{aligned} L_j(x) - \eta_A &\geq 0 && x \in H^1 \\ L_j(x) - \eta_B &\leq 0 && x \in H^0 \\ \eta_B < L_j(x) &< \eta_A && \text{Observe the next feature } x_{j+1} \end{aligned} \qquad \text{(III-11)}$$

It is well known that if we set

$$\eta_A = \frac{1 - \beta}{\alpha} \qquad \eta_B = \frac{\beta}{1 - \alpha} \qquad \text{(III-12)}$$

then the decision function of (III-11) has the property that among all sequential tests with the same specified α and β this SPRT will require the smallest number of features to reach a classification decision on the average.*

Computationally, the main problem in the use of SPRT is the recursive evaluation of the likelihood function. In general, for real time application one would like a formula of the type

$$L_{j+1}(x) = L_j(x) \times (\text{term involving } x_{j+1} \text{ only}) \qquad \text{(III-13)}$$

---

* The above statement is true to within the accuracy of the so-called "excess over boundary" represented by $L_j(x) - \eta_A$ or $L_j(x) - \eta_B$ [49].

or

$$\ln[L_{j+1}(x)] = \ln[L_j(x)] + \ln[x_{j+1} \text{ term}] \tag{III-13'}$$

This turns out to be possible if the x_i's belong to a fairly general class of gaussian sequences. In particular, let

$$\begin{aligned} x_j &= Hy_j + v_j \\ y_j &= \Phi y_{j-1} + w_{j-1} \end{aligned} \tag{III-14}$$

where v_j and w_j are independent white gaussian sequences with

$$\begin{aligned} E(v_j) &= 0 \qquad E(v_j v_i^T) = R_j \delta_{ij} \\ E(w_j) &= \bar{w}_j \qquad E[(w_j - \bar{w}_j)(w_i - \bar{w}_i)^T] = Q_j \delta_{ij} \end{aligned} \tag{III-15}$$

i.e. the features are noise corrupted linear combinations of the state of a vector gauss-markov sequence. Then a set of finite dimensional sufficient statistics of the features exist in the form of conditional mean and covariances of the state y, $(\hat{y}, P_y)$, that is

$$p(x_j/x_1, \ldots, x_{j-1}; H^i) \Longleftrightarrow p(x_j/\hat{y}, P_y; H^i) \tag{III-16}$$

These statistics can be recursively updated in terms of the "Kalman Filter" well known in control theory [26].

The relationship (III-16) and the observation

$$p(x_j, x_{j-1}, \ldots, x_1/H^i) = p(x_j/x_1, \ldots, x_{j-1}; H^i) p(x_1, \ldots, x_{j-1}/H^i) \tag{III-17}$$

immediately leads to Eq. (III-13). This powerful technique apparently has not been exploited to any great extent in the pattern classification literature.

Fu and his associates have studied various aspects of sequential methods as applied to pattern recognition. Chen [10] considered the reordering of the unobserved features so as to next observe the feature containing most significant information about the pattern. A straightforward SPRT is then applied to the features selected. For a SPRT with two parallel stopping boundaries as in Eq. (III-11) one can easily compute the average number of features required for a decision. This, however, does not guarantee that the decision process will terminate in every case. In practice it may not be possible to observe more than a finite number of features. Chien [11, 12] has suggested using a time varying stopping boundary to assure a termination in a finite time.

Consider the features $x_1, x_2, \ldots, x_n$ approximated by a continuous time function x(t). Our two hypotheses now involve determining x(t) as samples from one of the two stochastic processes H^o or H^1. Again the likelihood ratio can be formed and a continuous analog of a SPRT used for decision purposes:

$$L[x(t)] = \frac{p[x(t)/H^1]}{p[x(t)/H^o]} \tag{III-18}$$

The modified SPRT is stated by the following inequalities

$$\eta_B(t) < L[x(t)] < \eta_A(t) \tag{III-19}$$

where $\eta_B(t)$ and $\eta_A(t)$ are nondecreasing and nonincreasing (respectively) functions of time. The decision is made as to class H^o or H^1 when left or right inequality is violated. By making $\eta_A(t)$ and $\eta_B(t)$ functions of observation time it is possible to insure that a decision is reached in a finite time.

The expected time of reaching a decision and the probabilities of error, of course, will be different from the usual SPRT. But they may be calculated and controlled in advance.

One form for η_A and η_B may be the following

$$\ln \eta_A(t) = -a(1 - t/T)^{r_a}$$
$$\ln \eta_B(t) = b(1 - t/T)^{r_b} \qquad \text{(III-20)}$$

where $o \leq t \leq T$, $o < r_a$, $r_b \leq 1$, $a > o$, $b > o$.

T is the preassigned time at which the test is truncated. Let us denote the expected termination time for MSPRT* by $E'(t_T)$ and $E(t_T)$ for a standard SPRT, and by α' and β' the misclassification probabilities of a MSPRT. If α and β are very small and the boundaries of the Wald test and the MSPRT begin at the same point then

$$E'(t_T) \approx \frac{E(t_T)}{1 + \frac{r_b}{T} E(t_T)}$$

and

$$\alpha' = \alpha \left[1 + \frac{b r_b E(t_T)}{T + r_b E(t_T)} \right]$$

Therefore the expected time of termination is reduced and is controlled by the parameter r_b, while the error probability has increased. If it is desired that the same error probabilities be achieved in both tests, the boundaries of a MSPRT should begin at a higher value than those in Wald's SPRT.

---

* Modified Sequential Probability Ratio Test.

By arbitrarily assuming the form for the stopping boundary with undetermined parameters, e.g. Eq. (III-20), an optimal MSPRT w. r. t. the assumed structure can be designed. If we know the costs of continuing the observations and the cost of making a wrong decision, using available information at every instant the idea of dynamic programming can be used to arrive at the best stopping boundary using the standard idea of backward sweep and the principle of optimality. In a practical situation, this may result in excessively large amounts of data that cannot be handled. Chien [11] has suggested some procedures to reduce the total data to be handled at any stage.

Summarizing, we may say that case A is characterized by direct application of decision-theoretic ideas to pattern recognition. Because of the assumed availability of prior data, usually no iteration is involved in the determination of the decision function or separating surface.

IV. Case B - Data Type (i) and (iii) Are Given

When the functional form of conditional density function $p(x/H^i, \theta)$ is given but θ are unknown parameters, the obvious modification involves the use of the given sample patterns with known classification to estimate these parameters before performing hypothesis testing. The basic quantity of interest still is the <u>likelihood ratio</u> which is now defined as:

$$L(x) = \frac{p(x/x^1(1)\ldots x^1(n), H^i)}{p(x/x^o(1)\ldots x^o(n), H^o)} \triangleq \frac{p(x/\chi^1(n), H^1)}{p(x/\chi^o(n), H^o)} \qquad \text{(IV-1)}$$

We may write

$$p(x/\chi^i(n), H^i) = \int p(x/\theta, H^i)p(\theta/\chi^i(n), H^i)d\theta \qquad \text{(IV-2)}$$

Assuming the computation of (IV-2) is straightforward though it may be laborious, the determination of the conditional density $p(\theta/\chi^i(n), H^i)$ becomes the principal problem. We have by Bayes Rule,

$$p[\theta/\chi^i(n)] = \frac{p[x^i(n)/\theta, \chi^i(n-1)]p[\theta/\chi^i(n-1)]}{\int p[x^i(n)/\theta, \chi^i(n-1)]p[\theta/\chi^i(n-1)]d\theta}$$

$$= \frac{p(x^i(n)/\theta)}{\int p(x^i(n)/\theta)p[\theta/\chi^i(n-1)]d\theta}\, p[\theta/\chi^i(n-1)] \qquad \text{(IV-3)}$$

where the simplification in the second step comes about due to the assumed "conditional independence" of the sample patterns.*

We have also dropped the explicit dependence of $p[\theta/\chi^i(n)]$ in H^i for notational simplicity. It is understood that (IV-3) has to be carried out for each class.

Equation (IV-3) is a recursive computational procedure which is often referred to as "learning with teacher." The computational feasibility of (IV-3) depends critically on the existence of a fixed dimensional sufficient statistic for the relevant prior and posterior density functions. In other words, one would like to be able to compute recursively a vector, $\hat{\theta}^i_n$, with the property that

---

* By conditional independence we mean $p(x^i(n)/\theta) = p[x^i(n)/\theta, x^i(n-1)\ldots, x^i(1)]$.

$$p(\theta/\hat{\theta}^i_n) \Leftrightarrow p[\theta/\chi^i(n)] \tag{IV-4}$$

Then instead of doing recursion on functions which is the case for Eq. (IV-3), one is only concerned with updating a set of numbers, $\hat{\theta}^i_n$. Prior and posterior density functions which satisfy this requirement are called conjugate or reproducing pairs. They have been extensively studied by Raiffa and Schlaifer (1960)[40] and Spragin (1963)[45, 46]. It can be shown that in the limit of an infinite number of learning samples, the reproducing densities have the property that

$$\lim_{n\to\infty} \hat{\theta}_n \to \theta \tag{IV-5}$$

in some appropriate sense. Thus, this learning scheme used with any of the decision functions of case A is at least asymptotically optimal and in the limit produces results as good as if θ were known. In fact, if one interprets the α and β as average error probabilities with

$$\alpha \triangleq \int_{X^1} p(x/\theta, H^o)p(\theta/\chi^o(n), H^o)d\theta$$

$$\beta \triangleq \int_{X^o} p(x/\theta, H^1)p(\theta/\chi^1(n), H^i)d\theta \tag{IV-6}$$

then optimality for a finite number of learning samples can also be claimed. In general, however, the relationship between system performance and this learning scheme for finite samples is only qualitative and has not been investigated thoroughly. Putting it less precisely, we have the question: "Given the optimal decision function as a function of θ and the best estimate of θ, does the over-all optimal decision function simply involve the replacement of θ by its estimate?"

Special case of Gaussian $p(x/H^i, \theta)$

Consider the case where $p(x/H^i, \theta)$ is $N(\theta^i, \Sigma)$, Σ given, and let $p(\theta^i)$ be $N(\bar{\theta}^i, P)$. An easy way to treat this problem will be to consider

$$x^i = \theta^i + v \qquad \text{(IV-7)}$$

where v is $N(0, \Sigma)$.

Then $p(\theta^i/\chi^i(n)$ is gaussian with mean $\hat{\theta}^i_n$ and covariance P_n where

$$\hat{\theta}^i_n = \hat{\theta}^i_{n-1} + P_n(x^i(n) - \hat{\theta}^i_{n-1}) , \quad \hat{\theta}^i_o = \bar{\theta}^i \qquad \text{(IV-8)}$$

$$P_n = P_{n-1} - P_{n-1}(P_{n-1} + \Sigma)^{-1}P_{n-1} , \quad P_o = P \ . \qquad \text{(IV-9)}$$

This reduces $p(x/\chi^i_n, H^i)$ to a gaussian distribution with mean θ^i_n and covariance $P_n + \Sigma$. The numbers $\hat{\theta}^i_n$ and P_n constitute a set of finite dimensional sufficient statistics for $\chi^i(n)$. If Σ is the same for the two categories, we have the linear decision function as

$$f(x) = a^T x + a_o$$

$$a^T = (P_n + \Sigma)^{-1}(\hat{\theta}^1_n - \hat{\theta}^o_n) \ .$$

Note that equations (IV-8) and (IV-9) are a special case of the "Kalman Filter" mentioned in Eqs. (III-14) and (III-15) (with $\Phi = 0$ and $w_i = 0$). This was first worked out independently by Abramson and Braverman.[1]

If Σ is also unknown then the conjugate density is gauss-wishart (Keehn, 1963)[28]. If Σ_i are known but different or if θ^i are time varying and can be represented by a gauss-Markov process, then the theory of "Kalman Filter" can again be directly used to develop decision functions (or equivalently estimates for θ^i) that "tracks" the variations.

Special Case of Discrete Distribution

In the discussion so far, the learning of L(x), Eq. (IV-1-IV-3) and the determination of f(x), Eq. (III-2) are two separate problems. In certain simplified cases it is possible to devise a "learning" procedure for f(x) directly. Sklansky [44] has considered the classification of a sequence of independent binary signals transmitted over a noisy channel. Let x(j) be the channel outputs, we consider a decision function

$$f(x) = x - \eta \qquad \text{(IV-10)}$$

If the distribution of x as well as the choice of η values is discrete, then for a given procedure of changing η values after each wrong decision, the probabilities of η at the various permissible values form a markov chain. The property and convergence of this markov chain can be straightforwardly calculated once the transition probabilities (i.e. the learning procedures) are given. From this, the error probabilities of f(x) follows.

The performance of a scheme of this type depends to a large extent on the type of updating for η. Kaplan and Sklansky [27] have analyzed the properties of markov chains resulting from some typical learning procedures specified on an intuitive basis.

V. Case C - Data Type (i) and (iv) Are Given

In this case again the given set of sample patterns will be used to learn the parameters θ. But now the given learning samples are unclassified, bringing in additional uncertainty. Appropriately this type of learning is often called "Learning Without Teacher." Daly (1962)[17] suggested a scheme which works in this case but the

computation grows exponentially. Later Fralic suggested a bounded scheme[22, 21] which was further extended by Patrick and Hancock[37].

The basic ideas in section IV still apply here. We may rewrite Eq. (IV-3) as

$$p(\theta/\chi(n)) = \frac{p(x(n)/\theta, \chi(n-1)}{p(x(n)/\chi(n-1))}\ p(\theta/\chi(n-1))$$

$$= \frac{p\{x(n)/\theta, \chi(n-1), H^o\}P(H^o) + p\{x(n)/\theta, \chi(n-1), H^1\}P(H^1)}{p(x(n)/\chi(n-1), H^o)P(H^o) + p(x(n)/\chi(n-1), H^1)P(H^1)}$$

$$\cdot\ p(\theta/\chi(n-1)$$

The only difference occurs in the way we compute the ratio between "prior" and "posterior" density. The added term essentially represents a form of "hedging."

Heuristically, we can see the effect as follows. Consider the case where x(n) came from H^1. Now the multiplying factor on the right-hand side of Eq. (V-1) is of the form $\frac{A + C}{B + C}$ while with the additional knowledge about its class, it will only be $\frac{A}{B}$. For x(n) actually from H^1 we generally have $A > B$ and in this case $\frac{A}{B} > \frac{A + C}{B + C}$. This tends to indicate that the process of learning will be slower in the case of "Learning Without Teacher." Essentially we are paying for the uncertainty about the classification of learning samples in terms of slower learning. Viewed in this light, the difference between learning "With" or "Without" teacher is conceptually minimal. Another example illustrates this point. Consider the classification problem shown in Figure 1.

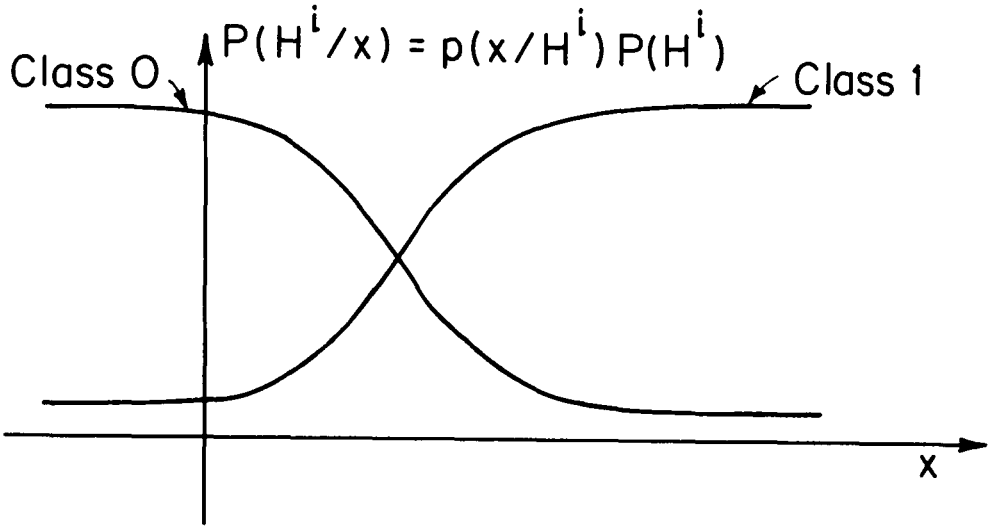


FIGURE 1

Suppose

(i) A set of samples $\chi(n)$ was taken and correct classification was attached to these.

(ii) A set of samples $\chi(n)$ was taken and classifications were assigned according to the probability $P(H^i/x)$.

It is clear that for $n \to \infty$ no difference in learning behavior can be observed from using samples from (i) or (ii) through the schemes of section IV.

Computationally, Eq. (V-1) is much more difficult than its counterpart, Eq. (IV-3). With the presence of the additional terms the "reproducing property" is lost.

Interpretation of Eq. (V-1) also requires some care. Since learning samples are unclassified, Eq. (V-1) cannot be carried out separately for each class in general. Let θ represent the unknown parameters in both classes, θ^1 and θ^o. If, in addition, we assume $p(\theta^1, \theta^o/\chi(n)) = p(\theta^1/\chi(n))\, p(\theta^o/\chi(n))$ then we can separate Eq. (V-1) into (V-2).

$$p(\theta^i/\chi(n)) = \frac{\sum_{j=0}^{1} p(x(n)/\theta^i, \chi(n-1), H^j)p(H^j)}{\sum_{j=0}^{1} p(x(n)/\chi(n-1), H^j)p(H^j)}\, p(\theta^i/\chi(n-1) \qquad \text{(V-2)}$$

$i = 0, 1$

Furthermore, we can usually unite $p(x(n)/\theta^i, \chi(n-1), H^j) = p(x(n)/\chi(n-1), H^j)$ $j \neq i$. Eq. (V-2) is essentially Fralick's scheme. Note if $p(\theta^i/\chi(n))$ is identical for $i = 0, 1$ and $P(H^o) = P(H^1)$ then no learning can take place.

In the work of Patrick and Hancock[37] the independence assumption of θ^0 and θ^1 is not made and computation must take place via the single Eq. (V-1) with the resultant added complexity. The only simplification is to note that $p(x(n)/\theta, \chi(n-1), H^j) = p(x(n)/\chi(n-1), H^j)$ for $j = 0, 1$.

The main advantage of learning without a teacher results from the fact that when actual processing of data is in progress (after initial learning from the sample patterns with classification) learning can still continue and eventually a machine learning this way may do much better than a machine which is trained only by initial learning with classified samples.

Very little computational result has been reported in the literature although our society seems to abound with real life examples of "learning without teacher" or even "learning in spite of the teacher."

VI. Case D - Data Type (iii) Given Only (Deterministic Methods)

The previous three sections had dealt with algorithms which require knowledge of the structural forms of the underlying distributions of the pattern classes. Criticism has often been raised that in practice information concerning data type (i) is seldom available. This has prompted development of algorithms for the construction of decision functions which do not require (i). Basically the idea is to find an f(x) which 'works well' at least on the given samples of known classification.

Two implicit assumptions of this approach are:

(a) A sufficient number of samples from both classes are available to constitute two representative groups.

(b) The characteristic problem (i.e. $x = \phi(z)$) has been solved using a sufficiently rich class of $\phi(z)$'s so that it is only necessary to examine the class of linear f(x) to solve the abstraction problem.

Assumption (ii) is often justified on the basis of the Weierstrass approximation theorem. However, this merely transfers the difficulty to the characterization problem since one is still faced with the problem of finding a class of complete $\phi(z)$'s which can efficiently represent the pattern. Furthermore, relatively little work has been done on rendering the adjectives "sufficient, representative, and efficient" quantity in the above assumptions. The works of Cover (1964)[15] Allais (1965)[4] and Watanabe (1965)[50] bear on this aspect of the problem. We shall discuss them separatly later.

Accepting (a) and (b), one can now restate the problem of abstraction more succinctly. Let there be a total of N patterns given (n_1 in class H^1, n_o in H^o, $n_o + n_1 = N$) and consider the linear decision function

$$f(x) = a^T x + a_o \qquad \text{(VI-1)}$$

The problem of determining a f(x) which classifies all the given patterns correctly is equivalent to the problem of finding a solution to the vector inequality

$$Aw > o \qquad \text{(VI-2)}$$

where

$$A = \begin{bmatrix} \underline{1} & -\underline{x}^{1T}(i) \\ -1 & -x^{oT}(i) \end{bmatrix} \begin{matrix} \} \\ \} \end{matrix} \begin{matrix} i = 1, \ldots, n_1 \text{ ; class 1 samples} \\ i = 1, \ldots, n_o \text{ ; class o samples} \end{matrix} \qquad \text{(VI-3)}$$

$$w = \frac{a_o}{a} \tag{VI-4}$$

A common procedure for solving linear inequalities is to transform it into an optimization problem the solution of which also guarantees a solution for (VI-2). For example, consider

$$\underset{w}{\text{Min}}\ J(w) \triangleq \underset{w}{\text{Min}}\ \| \ |Aw| - Aw \ \|^2 \tag{VI-5}$$

The solution for (VI-2), if it exists, must correspond to the minimum of (VI-5) which is zero. If we try a gradient descent procedure for minimizing (VI-5), then we are led immediately to

$$w(j+1) = w(j) + \rho \left.\frac{\partial J}{\partial w}\right|_{w=w(j)}$$

$$= w(j) + \rho A^T[\,|Aw(j)| - Aw(j)\,]\ ;\quad \rho > o \tag{VI-6}$$

or

$$w(j+1) = w(j) + \rho \sum_{i=1}^{N} x(i)[\,|x(i)^T w(j)| - x^T(i)w(j)\,] \tag{VI-6'}*$$

Algorithms of the type of Eq. (VI-6) are often referred to as "<u>many pattern adaptation</u>" in the sense that all given pattern samples are used in one iteration of the weighting vector "w". The corresponding <u>single pattern adaptation</u> of (VI-6) is

$$w(j+1) = w(j) + \rho\, x(j)(\,|x(j)^T w(j)| - x(j)^T w(j)) \tag{VI-7}$$

For $\rho = 2$, Eq. (VI-7) is simply the well-known peceptron algorithm (Novikoff (1962))[36] which was originally developed on the simple idea

---

* In (VI-6) we have abused our notation to let x(i) represent the vector $\begin{bmatrix} \pm 1 \\ x(i) \end{bmatrix}$.

of reward and punishment and which is known to converge in a finite number of steps.

The idea of viewing a learning algorithm as an iterative and deterministic optimization procedure for some criterion function can be used to interpret other algorithms and to discover new ones. In fact, our ability to create new algorithms is only limited by our ability to find meaningful new criteria. Table I identifies a set of algorithms as gradient procedures for a corresponding set of criterion functions.

The Generalization Question

One of the basic problems of the algorithms of this type is the question of generalization. In the absence of any probabilistic information, the only result along this line seems to be the important result of Cover (1964)[15]. Cover shows that in general the number of samples, N, must be at least equal to or larger than twice the number of attributes, m, for the algorithm of this case to yield meaningful results. If we allow ourselves the luxury of gaussian x_i's, then Allais demonstrates a more explicit relationship between N, m and the LMS algorithm of Table I.[4] Another interesting property of the LMS algorithm is pointed out by Groner [23]. It turns out that the $w \triangleq \{a_o, a\}$ which minimizes $\|Aw - \beta_o\|^2$ can also be expressed as

$$a = (\Sigma_s^1 + \Sigma_s^o)^{-1}(\mu_s^1 - \mu_s^o) \qquad \text{(VI-8)}$$

$$a_o = (\mu_s^1 + \mu_s^o)(\Sigma_s^1 + \Sigma_s^o)(\mu_s^1 + \mu_s^o)$$

where

TABLE I

Authors	Algorithms	Criterion J(w)
Widrow and Hoff [52]*	$w(j+1) = \rho A^T(Aw(j) - \beta_o)$	$J(w) = \|Aw - \beta_o\|^2$, $\beta_o^T = [1, 1, \ldots, 1]$
Agmon - Mays [2, 30]	$w(j+1) = w(j) + \rho A^T[\|Aw(j) - \beta_o\| - (Aw(j) - \beta_o)]$	$J(w) = \|(Aw - \beta_o) - \|Aw - \beta_o\|\|^2$, $\beta_o^T = [1, 1, \ldots, 1]$
Wong - Eisenberg [51]**	$Aw(j+1) = Aw(j) + \rho A(A^TA)^{-1}A^T[\beta_o - \mathrm{sgn}(Aw(j))]$	
Ho - Kashyap [25]	$w(j+1) = w(j) + \rho SA^T[Aw(j) - \beta(j)]$ $\beta(j+1) = \beta(j) + [(Aw(j) - \beta(j) + \|Aw(j) - \beta(j)\|]$ where ρ and S are chosen to insure $[\rho^2 SA^TAS - 2\rho S] < o$	$J(w, \beta) = \|Aw - \beta\|^2$, $\beta > 0$

* Solution of this algorithm is not equivalent to the solution of $Aw > o$.

** This algorithm is different from the Ho - Kashyap algorithm with $S = (A^TA)^{-1}$ only in the sense that β_o is constant here.

μ_s^i = sample mean of class H^i

Σ_s^i = sample covariance of class H^i .

Furthermore, the "a" in Eq. (VI-8) also maximizes the Maholanobis distance criterion[5]

$$J = \frac{[a^T(\mu_s^1 - \mu_s^o)]^2}{a^T(\Sigma_s^1 + \Sigma_s^o)a} \qquad \text{(VI-9)}$$

which has the simple interpretation of maximizing interclass distance and minimizing total dispersion of the projections of the patterns onto the decision surface f(x).

Eq. (VI-8) can be further generalized. Peterson and Mattson[38] have shown that for a linear decision function of the type of Eq. (VI-1) and a criterion function J which depends only on sample means and covariances of the two classes, the optimal "a" is given via

$$a = (k_1 \Sigma_s^1 + k_o \Sigma_s^o)^{-1}(\mu_s^1 - \mu_s^o) \qquad \text{(VI-10)}$$

where k_o and k_1 are constants that can be determined.*

While these connections of the linear decision function and the statistical parameters of the sample patterns are interesting, they do not completely answer the generalization question in terms of the error probabilities α and β of (III-3) and (III-4).

A recent important result of Cover and Hart (1967)[16] is an exception to this point. They show that if we classify a sample by the classification of its nearest (according to some distance measure)

---

* The validity of (VI-10) is, of course, still good if we replace sample means and covariances by true mean and covariances.

neighboring sample of known classification, the error probability of such a decision is bounded from above by twice the optimal Bayes error probabilities of section III(ii) when all the underlying probabilities are known. This Nearest Neighbor decision function is originally due to Fix and Hodges[53].

Two other statistical techniques commonly used in data analysis called "jacknife" and "leaving-one-out"[33] may be useful to shed further light on this question. This technique consists of successively solving a series of optimization problems each time leaving out a different sample pattern. Variation in the solutions of these problems will then indirectly provide a quantitative answer to the adequateness of assumption (i). This approach apparently has not been exploited in the usual pattern classification literature.

The Characterization Problem

Although we conveniently avoided the question of how to choose a mapping $\phi : z \rightarrow x$ the question is nevertheless an important one. There do not seem to be many generally applicable schemes which possess noteworthy properties that are independent of the particular type of recognition problems in question. An important result due to Watanabe (1965)[50] does, however, fill the requirement. Consider the (sample) covariance matrices of each class, Σ_{z_1} and Σ_{z_o} and the linear combination

$$\Sigma = P(H^1)\Sigma_{z_1} + P(H^o)\Sigma_{z_o} \qquad \text{(VI-11)}$$

Let the dim. (z) = p which is usually very large compared to the desired dimension for x, m. Let the vectors $t_1, t_2, \ldots, t_p$ be the normalized

eigenvectors of Σ ordered according to $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \ldots \geq \lambda_p(\Sigma)$.

The t_i's form a basis in p space.* We may write

$$z = \sum_{i=1}^{p} x_i t_i \qquad \text{with} \qquad x_i = z^T t_i \tag{VI-12}$$

The magnitude of x_i or more accurately $P(H^1)(x_i^1)^2 + P(H^o)(x_i^o)^2 \triangleq e_i$ can be considered as a good measure of the extent to which the coordinate vector t_i is useful in representing the members of the two classes. It turns out in this set up that the following properties are true:

(i)
$$E\left\{P(H^1)(z^1 - \sum_{i=1}^{m} x_i t_i)^2 + P(H^o)(z^o - \sum_{i=1}^{m} x_i t_i)^2\right\} =$$

$$\underset{s_i}{\text{Min}}\, E\left\{P(H^1)(z^1 - \sum_{i=1}^{m} x_i s_i)^2 + P(H^o)(z^o - \sum_{i=1}^{m} x_i s_i)^2\right\} \quad \text{for all m.} \tag{VI-13}$$

i.e. the t_1 coordinate system has the least square approximation property.

(ii) The t_i coordinate system minimizes the entropy function

$$S(t_i) \triangleq -\sum_{i=1}^{p} e_i \ln e_i \tag{VI-14}$$

among all possible coordinate systems, where e_i's are as defined above. (i) and (ii) imply that there exists a natural (according to (VI-13) and (VI-14)) characterization of the attributes in terms of the large eigenvectors of the composite covariance matrix of the problem.

---

* They constitute the Karhunen-Loeve coordinate system.

This phenomenon is related to the method of factor analysis and which we shall encounter again in section VIII.

Extension to Nonlinear f(x)

Once the approach for the linear case is clear, the extension to the nonlinear case is conceptually straightforward. Instead of linear inequalities, we deal with nonlinear inequalities or piecewise linear inequalities. Various established or ad hoc techniques in non-linear programming can be brought to bear on the abstraction problem[42].

Arkadev and Braverman[7] have suggested an algorithm which tries to arrive at a piecewise linear f(x) from the given set of classified x's. The algorithm proceeds to find the best hyperplane to separate the given samples. As more samples come and this plane fails to classify them, more hyperplanes are connected to it until all given samples are correctly classified. Finally all redundant portions of the planes are deleted giving a piecewise linear f(x).

VII. Case E. Data Type (iii) Only Given (Stochastic Methods)

The algorithms described in section VI with the exception of the NN algorithm, despite their simplicity and practical usefulness, suffer one general drawback in terms of relating the decision functions obtained to a quantitative evaluation of its generalization capabilities. Since the latter question is best answered in terms of error probabilities, or equivalently the knowledge of $P(H^i/x)$, some probabilistic structure will have to be put back into the formulation of the problem, implicitly or explicitly.

One approach to this problem is to consider that $P(H^i/x)$ as a function of x can be expanded in a series. Let us consider

$$f(x) \triangleq P(H^1/x) - P(H^o/x) \triangleq 2P(H^1/x) - 1$$

$$\triangleq \sum_{j=o}^{\infty} a_j \phi_j(x) \qquad \text{(VII-1)}$$

where $\phi_i(x)$ is some class of complete (possibly orthonormal) function which one conveniently assumes to be given as a result of solving the characterization problem. For every given sample pattern, there corresponds a $\phi[x(i)]$ or $\phi(i)$. The problem is then simply reduced to the determination of the parameters, a_j, of a function f(x) when the values of the function are measured at randomly selected points. This is essentially the approach taken by Aizerman, Baverman and Rozonoer[3], Tsypkin[48], Blaydon and Ho[8], Kashyap and Blaydon[29], Patterson, Wagner and Womack[39], and Nicolic and Fu[35]*. Define a classification variable $\zeta[x(i)]$ by

$$\zeta[x(i)] = \begin{cases} 1 & x(i) \in H^1 \\ -1 & x(i) \in H^o \end{cases} \qquad \text{(VII-2)}$$

One may visualize $\zeta(i)$ as a noisy measurement of the value of the function f(x) at the sample point (pattern) $\zeta(i)$.

$$\zeta(i) \triangleq f[\phi(i)] + v(i) \qquad \text{(VII-3)}$$

where v(i) are independent random variables with

$$E[v(i)] = [1 - f(i)]P[H^1/x(i)] + [-1 - f(i)][1 - P(H^1/x(i))]$$

$$= 2[1 - P(H^1/x(i)]P(H^1/x(i)) - 2P(H^1/x(i))(1 - P(H^1/x(i))] = 0$$

$$\text{(VII-4)}$$

---

* These methods are also called "Potential Function" methods[7], in the Russian literature.

Now consider the minimization of the regression function

$$J = \min_{a} E\{ \| \zeta - a^T \phi \|^2 \} \tag{VII-5}$$

which in view of Eqs. (VII-3-VII-4) can be shown to be equivalent to

$$J^1 = \min_{a} E\{ \| f(x) - a^T \phi \|^2 \} \tag{VII-6}$$

Thus if one finds a finite dimensional a which minimizes J in Eq. (VII-5) then one has also found the optimal mean square approximation to f(x). A well known method for minimization of regression functions is via stochastic approximation using the given noisy sample values of the function. We have

$$a(i+1) = a(i) + \rho(i)\, \phi(i)[\zeta(i) - a^T(i)\, \phi(i)] \tag{VII-7}$$

with

$$\sum_{i=1}^{\infty} \rho(i) = \infty \; ; \quad \sum_{i=1}^{\infty} \rho^2(i) < \infty \tag{VII-8}$$

With mild assumptions on ϕ, the algorithm of Eq. (VII-7) is known to converge w.p.1. to a^* where

$$a^* = \arg \min \; E\{ \| \zeta - a^T \phi \|^2 \} \tag{VII-9}$$

Hence $\qquad = \arg \min \; E\{ \| f(x) - a^T \phi \|^2 \}$

This constitutes a learning scheme with teacher which is asymptotically optimal in a mean square sense w.r.t. the classification probabilities. Another way of visualizing (VII-7) is to note that

$$\phi(i)[\zeta(i) - a^T(i)\phi(i)] = \left. \frac{\partial \| \zeta - a^T \phi \|^2}{\partial a} \right|_{\phi(i)}$$

and (VII-7) is simply the stochastic analog of the gradient method for minimization. In fact, if one considers instead

$$\alpha(i+1) = \alpha(i) + \rho(i)\, S\phi(i)[\zeta(i) - \alpha^T(i)\phi(i)] \qquad \text{(VII-7')}$$

where

$$S \triangleq [E(\phi\phi^T)]^{-1} \qquad \text{(VII-10)}$$

one has the analog of the second order descent method which generally converges faster. Since S in Eq. (VII-10) is not given in general, one may substitute instead

$$S(i) = [\frac{1}{i} \sum_{j=1}^{i} \phi(j)\phi(j)^T]^{-1} \triangleq i\,P(i) \qquad \text{(VII-11)}$$

Not surprisingly, the recursive computation of P(i) is governed by

$$P(i+1) = P(i) - P(i)\phi(i)\{\phi(i)^T P(i)\phi(i) + 1\}^{-1}\phi(i)^T P(i)\ ;\ P(1) = I \qquad \text{(VII-12)}$$

which is a special case of Eq. (IV-9). Furthermore, $\alpha(i+1)$ from Eq. (VII-7) has the property that

$$\alpha(i+1) = \text{arg. min.} \sum_{j=1}^{i+1} [\zeta(j) - \alpha^T\phi(j)]^2$$

In other words, it is also the solution of the LMS algorithm of section VI. Using the method of stochastic approximation, one can show[8] that Eq. (VII-7') also converges w.p.1. to α^*, thus, furnishing additional rationalization for the LMS algorithm.

The choice of the base function set ϕ has been so far left open. A particular approach to the problem has been suggested by Brick[9]. Instead of expending f(x) in a series, $p(x/H^i)$ may be expended in a series of orthonormal functions, the normalized Hermite functions. Brick has shown that these coefficients appear as some ensemble average which, given certain a priori information, can be precomputed.

In case of ergodic processes, however, these ensamble averages can be replaced by time averages and can be easily determined experimentally given a set of classified samples for initial learning. Furthermore, if the system parameters are known to change gradually a bootstrap updating of these coefficients is possible during the run, improving continuously on the "Learned" coefficients.

The main advantage of this scheme is the possibility of implementation in circuit form for ergodic processes, where the coefficients appear only as amplifier gains which can be easily preadjusted or changed automatically. The number of terms to be used in the expansion depends, however, on the complexity of the form of $p(x/H^i)$. In practical situations the implementation may not be feasible for any moderately complex system.

VIII. Case F - Data Type (iv) Given Only

This is the extreme case in pattern classification where minimal information is available for the design of f(x). Not much has been reported in the literature about the approaches for this case which are often heuristic or experimental, justified only by the fact that they "work" in some sense according to the author. Analytically, the problem can be resolved in either one of two ways:

(a) Reintroduce, explicitly or implicitly, some criterion of separation onto the set of unclassified sample patterns. This is used in conjunction with the same algorithms of the case D in a bootstrap fashion, i.e. one uses the result of classification at one iteration to produce the "classified" learning sample for the learning cycle. For example a bootstrap algorithm results if we try to rewrite Eq. (VII-5) as

$$J = \underset{a}{\text{Min}}\ E\{\|a^T\phi - \text{Sgn}(a^T\phi)\|^2\} \qquad \text{(VIII-3)}$$

Miller[31] has examined this criterion function for a linear f(x) i. e. where ϕ is only x. For this case J becomes

$$J' = \underset{a}{\text{Min}}\ E\{\|a^T x - \text{Sgn}(a^T x)\|^2\} \qquad \text{(VIII-3')}$$

and an algorithm parallel to Eq. (VII-7) may be given as

$$a(k+1) = a(k) - 2\rho(k)[a^T(k)x(k) - \text{Sgn}\{a^T(k)x(k)\}]x(k) \qquad \text{(VIII-4)}$$

with

$$\sum_{k=1}^{\infty} \rho(k) = \infty \quad ; \quad \sum_{k=1}^{\infty} \rho^2(k) < \infty$$

Miller found that the J' surface, in general, has saddle points and local minimums. In the case of gaussian distributions with the same covariance matrix and $\mu_1 = -\mu_o$, the J' surface has two local minimums with the same minimum value and the algorithm of Eq. (VIII-4) converges to one of the two minima w. p. 1. However, in this case the value of a at one minimum is the negative of the value at the other. As the decision takes place according to $a^T x \gtrless 0$ the decision surface using either value merely results in the relabelling of the classes.

It should be pointed out that if this externally imposed criterion of separation happens to resemble the natural criterion of separation then all is well. For example, consider the two-dimensional examples shown in Fig. 2 where the actual identity of the sample points is unknown to the classifier. One can nevertheless require that a separating plane (linear decision function) be constructed with

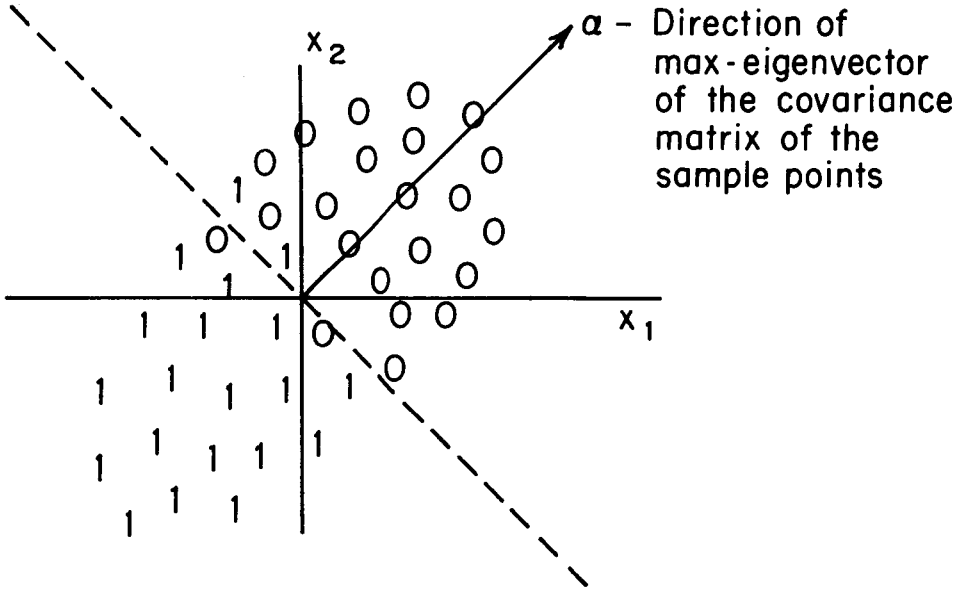

α - Direction of max-eigenvector of the covariance matrix of the sample points
x2
x1

FIGURE 2

the normal "α" coincide with the direction of the maximal eigenvector of the covariance matrix of the sample points, i.e.

$$\alpha \text{ parallel to the max eigenvector of } (A^T A) \qquad \text{(VIII-5)}$$

where

$$A = \begin{bmatrix} - \underline{x}^T(1) - \\ \vdots \\ - \underline{x}^T(N) - \end{bmatrix} \qquad \text{(VIII-6)}$$

If this eigenvector direction is determined via the usual iterative power method then one has derived a "learning" scheme which "works." However, it is equally obvious that had the two classes been distributed as in Fig. 3 the learning scheme would have failed. The above approach actually formed the essence of a very successful bootstrap self-correction scheme by Nagy and Shelton[34] for character recognition and is closely related to the method of principal component in factor analysis.

A slightly difference approach in introducing a criterion has been taken by Rogers and Tanimoto[4]. A distance measure d_{ij} for x_i and x_j may be defined as

$$d_{ij} = g(x_i, x_j) \qquad \text{(VIII-7)}$$

Assuming there are only two classes, we define a homogeneity function for each class. The function has d_{ij}, the interpair distances of the various sample patterns as its arguments. The assignment of a given sample pattern to a particular class changes the value of the homogeneity function of that class. The criterion of classification is that the two homogeneity functions have minimal

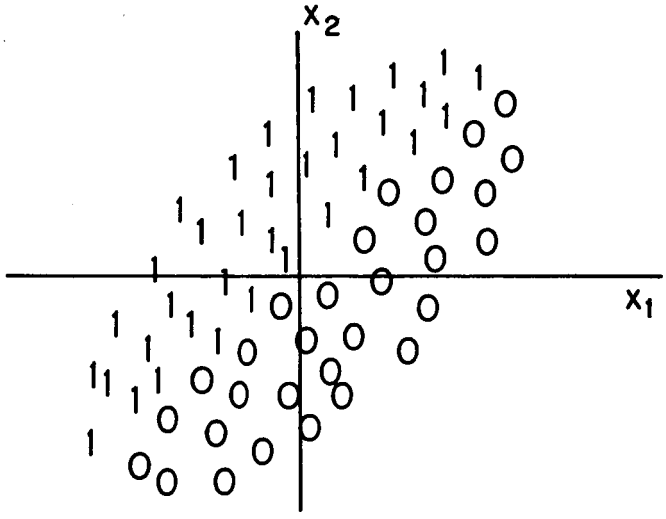


FIGURE 3

difference in their value. Analytically we may visualize this as follows:

Let

$$u_1(d_{12}, d_{13}, \ldots\ldots\ldots, d_{n-1,n}; \eta)$$

$$u_0(d_{12}, d_{13}, \ldots\ldots\ldots, d_{n-1,n}; \eta)$$

be the homogeneity functions involved. η is a parameter which assigns the usage of a particular d_{ij} to the evaluation of u_0 or u_1*. Classification then involves the choice of η such that the difference $u_1 - u_0$ is minimized.

This scheme can be very easily generalized to a multiclass case where the number of classes is not known a priori. In this light we shall encounter it again in the next section.

(b) Attempt a more or less brute force computation of the learning equation (V-1) in section V. The pertinent probability density functions are approximated by histograms using the given data. Patrick and Hancock claim convergence of this computational procedure[37]. However, little or no actual experiences have been reported. Due to the lack of sufficient statistics and the large amount of data usually required for histograms to yield good approximations, the feasibility of such a scheme remains to be demonstrated.

---

* For example: Roger and Tanimoto will successively include more and more d_{ij} in the evaluation of u_0 or u_1 until they exceed a value. The assumption here is that each class should be homogeneously similar. If for a given η, $u_1 >> u_0$, then some member of class 1 should be class 0 and vice versa.

IX. Multiclass Problem

So far we have concerned ourselves only with two-class problems. A few words about multiclass problems are in order.

In a multiclass problem we have to decide to which of m classes H^1, H^2, H^3, . . ., H^m the given pattern vector x belongs. For problems having a probabilistic structure i.e. from case A, B, C and E, the extension to a multiclass case is straightforward. Instead of considering the likelihood ratio formed by the two conditional probabilities and testing it against a threshold for a two-class case, we may directly consider $p(x/H^i)$ or $P(H^i/x)$ and pick the largest of these. This procedure is optimal in the sense that it minimizes the error probability.

More generally, we may define a set of decision functions $f^i(x)$ and take decisions as x belonging to H^i if

$$f^i(x) > f^j(x) \qquad \text{for all } j \neq i \tag{IX-1}$$

for the problems of case A, B, C and E

$$f^i(x) = p(x/H^i) \qquad \text{or} \qquad P(H^i/x) \tag{IX-2}$$

In the case of deterministic algorithms (nonprobabilistic structure) the extension to multi-class problems is not as natural. A multiclass problem is usually reduced to a collection of two-class problems. This reduction depends on the separability which exists in the multiclass problem and may be of three types.

(i) Each class may be separable from all the rest by a single decision surface. Then we may take the decision according to

$$f^i(x) \begin{cases} > o & \text{if } x \in H^i \\ < o & \text{otherwise} \end{cases}$$

This reduces the multiclass problem to m - 1 two-class problems

(ii) Each class may be separable from each other class. Now we have $\frac{m(m-1)}{2}$ two-class problems and as many decision functions such that

$$f^{ij}(x) \begin{cases} > o & \text{if } x \in H^i \\ < o & \text{if } x \in H^j \end{cases} \quad \text{given x belongs to } H^i \text{ or } H^j \tag{XI-3}$$

The unknown x is classified as H^i only if*

$$f^{ij}(x) > o \quad \text{for all } j \neq i$$

(iii) There exist m $f^i(x)$ such that x belongs to H^i only if

$$f^i(x) > f^j(x) \quad \text{for all } j \neq i \tag{XI-4}$$

Note that this is a special case of (ii) as we may define

$$f^{ij}(x) = f^i(x) - f^j(x) \quad \text{for all } j \neq i \tag{XI-5}$$

In cases A, B, C and E, as we noted earlier, $f^i(x)$ of Eq. (XI-4) are the conditional probabilities. For deterministic case D a criterion of separation is introduced. For example one may consider the inverse of distance criterion which is the extension of the Nearest Neighbor approach to this case.

The way we may try to reduce a multiclass problem to a set of two-class problems depends on the individual problem. It should be pointed out that if we are not restricting $f^i(x)$ to be of a particular form (linear or quadratic etc) then the distinction of these three types is artificial as we can always find suitable $f^i(x)$ to use with any of the desired three types.

---

* Some irrelevant results will be obtained in the process from the cases where x coming from H^k, $k \neq i, j$ is being tried.

Though we have presented all the algorithms as they are applicable to two-class problems, often the authors have extended it to the multiclass case. The number of classes is assumed known in all these schemes. One exception is the paper by Rogers and Tanimoto[41] mentioned under case F. The procedure they use is independent of the knowledge of the number of classes. By computing an auxiliary index of "typicality" they permit the procedure to adjust itself to produce automatically the number of classes in order to satisfy a homogeneity criterion.

X. Conclusion

Based on the above survey and analysis, a few remarks (perhaps controversial) seem in order:

(i) Roughly speaking, classification algorithms can be broadly divided into two groups; probabilistically or nonprobabilistically based. The former group, comprising cases A-C, enjoys the obvious advantage of being easy to assess the generalization ability of the results. On the other hand, it is often difficult to justify the availability of the required input data. The latter group representing cases D-F is just the reverse. By being more realistic on input data requirements it made precise quantitative evaluation of performance much more involved and difficult. Although it seems reasonable to assume that as our analytical ability advances this difficulty will gradually ease. The main "learning" tool for the first group is the recursive application of the Bayes Rule while that of the latter is the iterative solution of an optimization criterion. Both techniques are fundamental to stochastic and deterministic control theory. It

is expected that further cross fertilization will take place between these two fields.

(ii) "Characterization" remains to be a major open problem.

(iii) Relatively little experimentation with these algorithms have been carried out with real life classification problems compared to the number of proposed approaches. This is not so much a general criticism of the papers but a comment on the difficulties of obtaining real data. An often overlooked and unappreciated problem is that of converting or generating enough samples of x(i) from original data in real problems. Enormous data processing time or specially designed data processing machines are needed. In fact, it is the authors' belief that this often represents the major cost of a pattern classification project. Once this is done, the solution of the abstraction problem tends to be straightforward.

In this paper we have attempted a classification of the various pattern classification techniques that have been reported in the literature. The purpose has been to try to lay bare the underlying statistical and mathematical principles used in the development of these algorithms. Only when such a classification is complete then can meaningful comparisons among the numerous approaches be made. Furthermore, deficiencies as well as advantages of the various schemes hopefully can be made obvious and progress of the field as a whole can be sped up.

A complete coverage of the literature in this field is neither possible nor desirable. It is nevertheless believed that our classification is reasonably complete and workable, and future approaches

can be fitted into this framework, i.e. our scheme possesses the generalization property. There are many other papers which have appeared in literature. We leave it as an excercise for the reader to find their case classifications. The authors welcome additions to their bibliography for the various cases.

References

1. Abramson, N. & Braverman, D., "Learning to Recognize Patterns in a Random Environment," IRE Trans. on Info. Th., Vol. I. T. -8, pp. 58-63, September 1962.

2. Agmon, S., "The Relaxation Method for Linear Inequalities," Canad. J. Math., 6 (1956), pp. 382-392.

3. Aizerman, M. A., Braverman, E. M., & Rozonoer, L. I., "Method of Potential Functions in the problem of Restoration of Functional Converter Characteristic by Means of Points Observed Randomly," Automation and Remote Control, Vol. 25, No. 12, December 1964.

4. Allais, D. C., "Selection of Measurement for Prediction," Stanford Report No. 6103-9, November 1964.

5. Anderson, T. W., "An Introduction to Multivariate Statistical Analysis," Wiley, New York, 1958.

6. Anderson, T. W. & Bahadur, R., "Classification into Two Multi-Variate Normal Distributions With Different Covariance Matrices," Annals Math. Statistics, Vol. 33, 1962, pp. 422-431.

7. Arkadev & Braverman, "Computers and Pattern Recognition," Thompson, Washington, 1967.

8. Blaydon, C. C. & Ho, Y. C., "Recursive Algorithms for Pattern Classification," Proceedings of the National Electronics Conference, 1966.

9. Brick, D. B., "Wiener's Nonlinear Expansion Procedure Applied to Cybernetics Problems," IEEE Trans. on System Science and Cybernetics, Vol. SSC-1, pp. 67-74, November 1966.

10. Chen and Fu, "Sequential Decisions, Pattern Recognition and Machine Learning," TR-EE65-6, Purdue University, April 1965.

11. Chien, Y. T. & Fu, K. S., "On the Finite Stopping Rules and Nonparameter Techniques in a Feature Ordered Sequential Recognition System," Purdue University Report TR-3366-16, October 1966.

12. Chien, Y. T. & Fu, K. S., "Sequential Recognition Using a Nonparameter Ranking Procedure," IEEE Trans. on Info. Th., Vol. IT-13-, No. 3, July 1967, pp. 484-492.

13. Cooper, P. W., "Quadratic Discriminant Functions in Pattern Recognition," IEEE Trans. on Info. Th., Vol. IT-11, pp. 313-315, April 1965.

14. Cooper, P. W., "Hyperplanes, Hyperspheres and Hyperspadrices as Decision Boundaries," Computer and Information Science, Spartan Books Washington, 1964, pp. 111-139.

15. Cover, T. M., "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," IEEE Trans. On Electronic Computers, June 1965, Stanford Report 6107-1, May 1964.

16. Cover, T. M. & Hart, D. E., "Nearest Neighbor Pattern Classification," IEEE Trans. on Information Th., Vol. IT-13, No. 1, January 1967, pp. 21-26.

17. Daly, R. F., "The Adaptive Binary -- Detection Problem on the Real Line," Stanford Report TR 2003-3, February 1962.

18. Duda & Fossum, "Pattern Classification by Iteratively Determined Linear and Piecewise Linear Discriminent Functions," IEEE Trans. on Electronic Computers EC-15, pp. 220-232, April 1966.

19. Dreyfus, H. L., "Alchemy and Artificial Intelligence," Rand Report P-3244, December 1965.

20. Flanagan, J. L., "Speech Analysis, Synthesis and Perception," Academic Press, New York 1965.

21. Fralick, S. C., "The Synthesis of Machines Which Learn Without a Teacher," Stanford Report 6103-8, April 1964.

22. Fralick, S. C., "Learning to Recognize Patterns Without a Teacher," IEEE Trans. on Info. Th., Vol. IT-13, No. 1, pp. 57-65, January 1967. Stanford System Theory Lab. Report No. 6103-10, March 1965.

23. Groner, G. F., "Statistical Analysis of Adaptive Linear Classifiers," Stanford Electronic Laboratories Tech. Report No. 6761, April 1964.

24. Ho, Y. C. & Agrawala, A. K., "On the Self-Learning Scheme of Nagy and Shelton," Proc. of IEEE, Vol. 55, No. 10, pp. 1764-1765, October 1967.

25. Ho, Y. C. & Kashyap, R. L., "A Class of Iterative Procedures for Linear Inequalities," J. SIAM Control, Vol. 4, No. 1, 1966, pp. 112-115.

26. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans., ASME, J. Basic Engineering Series D, Vol. 82, pp. 35-45, 1960.

27. Kaplan, K. R. & Sklansky, J., "Analysis of Markov Chain Models of Adaptive Processes," AMRL-TR-65-3, Aerospace Medical Research Labs., January 1965.

28. Keehn, D. G., "A Note on Learning for Gaussian Properties," IEEE Trans. on Info. Th., Vol. IT-11, pp. 126-132, January 1965.

29. Kashyap, R. L. & Blaydon, C. C., "Recovery of Functions from Noisy Measurements Taken at Randomly Selected Points and Its Application to Pattern Classification," Proc. of IEEE, Vol. 54, No. 8, pp. 1127-1128, August 1966.

30. Mays, C. H., "Effect of Adaptation Parameters on Convergence Time and Tolerance for Adaptive Threshold Elements," IEEE Trans. on Electronic Computers, Vol. EC-13, 1964, pp. 465-468.

31. Miller, W. C., "A Modified Mean Square Criterion for use in Unsupervised Learning," Stanford Report, Center for Systems Research, TR No. 6778-2, August 1967.

32. Minsky, M., "Steps Towards Artificial Intelligence," Proc. IRE, January 1961, pp. 8-30.

33. Mosteller, F. & Tukey, J. W., "Data Analysis, Including Statistics," Handbook of Social Psychology, Editors: Gardner, Lindsay, Elliot and Aronsen; Addison Wesley, 1968.

34. Nagy & Shelton, "Self-corrective Character Recognition System," IEEE Trans. on Info. Theory, Vol. IT-12, No. 2, April 1966, pp. 215-222.

35. Nikolic, Z. J. & Fu, K. S., "A Mathematical Model of Learning in an Unknown Random Environment," Proceedings of the National Electronics Conference, Vol. 22, October 1966, pp. 607-612.

36. Novikoff, A., "On Convergence Proofs for Perceptrons," Proc. 1962 Symp. on Mathematical Theory of Automata, Polytechnic Inst. of Brooklyn, pp. 615-622.

37. Patrick, E. A. & Hancock, J. C., "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. on Info. Theory, Vol. IT-12, No. 3, July 1966, pp. 362-372.

38. Peterson, D. W. & Mattson, R. L., "A Method of Finding Linear Discriminant Functions for a Class of Performance Criteria," IEEE Trans. on Info. Th., Vol. IT-12, No. 3, July 1966.

39. Patterson, J. D., Wagner, T. J. & Womack, B. F., "A Performance Criterion for Adaptive Pattern Classification System," IEEE Trans. on Auto. Control, Vol. AC-12.

40. Raiffa, H. & Schlaifer, R., "Applied Statistical Decision Theory," Harvard Business School, Boston, 1961.

41. Rogers, D. J. & Tanimoto, T. T., "A Computer Program for Classifying Plants," Science, Vol. 132, pp. 1115-1118, October 1960.

42. Rosen, J. B., "Pattern Separation By Convex Programming," Journal of Mathematical Analysis and Application, Vol. 10, 1965, pp. 123-134.

43. Selin, "Detection Theory," Princeton University Press, 1965.

44. Sklansky, J., "Threshold Training of Two-Mode Signal Detection," IEEE Trans. on Info. Th., July 1965, pp. 353-362.

45. Spragins, J., "A Note on the Iterative Application of Bayes' Rule," IEEE Trans. on Info. Th., Vol. IT-11, pp. 544-549, October 1965.

46. Spragins, J., "Learning Without a Teacher," IEEE Trans. on Info. Th., Vol. IT-12, No. 2, pp. 223-229, April 1966.

47. Spragins, J. D., "Reproducing Distributions for Machine Learning," Stanford Report, 6103-7, November 1963.

48. Tsypkin, Y. Z., "Adaptation, Training and Self-Organization in Automatic Systems," Trans. from Automatika I. Telemeshanika, Vol. 27, No. 1, pp. 23-61, January 1966.

49. Wald, "Sequential Analysis," Wiley, New York 1967.

50. Watanabe, S., "Karhunen-Loeve Expension and Factor Analysis, Theoretical Remarks and Applications," Information Theory, Statistical Decision Functions, Random Processes Trans. of the Fourth Prague Conference, 1965, pp. 635-660.

51. Wong, E. & Eisenberg, E. "Iterative Synthesis of Threshold Functions," J. Math. Analysis Applications, Vol. 11, 1965, pp. 226-235.

52. Widrow, B. & Hoff, M. E., "Adaptive Switching Circuits," Stanford Electronics Laboratories Technical Report, TR No. 1553-1, June 1960.

53. Fix. E. & Hodges, J. L., "Discriminatory Analysis, Non-parametric Discrimination," USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004 Report 4, Contract AF 41-(128) 31, February 1951.

Joint Services Electronics Program

N00014-67-A-0298-0006, 0005, and 0008

Academy Library (DFSLB)
U. S. Air Force Academy
Colorado Springs, Colorado 80912

AEDC (ARO, INC)
Attn: Library/Documents
Arnold AFB, Tenn. 37389

Aeronautics Library
Graduate Aeronautical Laboratories
California Institute of Technology
1201 E. California Blvd.
Pasadena, California 91109

Aerospace Corporation
P. O. Box 95085
Los Angeles, Calif. 90045
Attn: Library Acquisitions Group

Airborne Instruments Laboratory
Deerpark, New York 11729

AFAL (AVTE/R. D. Larson)
Wright-Paterson AFB
Ohio 45433

AFCRL (CRMXLR)
AFCRL Research Library, Stop 29
L. G. Hanscom Field
Bedford, Mass. 01731

AFETR (ETLIG - 1)
STINFO Officer (for library)
Patrick AFB, Florida 32925

AFETR Technical Library
(ETV, MU-135)
Patrick AFB, Florida 32925

AFFTC (FRBPP-2)
Technical Library
Edwards AFB, Calif. 93523

AFGC (PHBPS-12)
Eglin AFB
Florida 32542

AFL (AFIY)
Wright-Paterson AFB
Ohio 45433

AULIT-9663
Maxwell AFB
Alabama 36112

Mr. Henry L. Bachman
Assistant Chief Engineer
Wheeler Laboratories
122 Cuttermill Road
Great Neck, N. Y. 11021

Bendix Pacific Division
11600 Sherman Way
North Hollywood, Calif. 91605

Colonel A. D. Blue
RTD (RTTL)
Bolling AFB
Washington, D. C. 20332

California Institute of Technology
Pasadena, California 91109
Attn: Documents Library

Carnegie Institute of Technology
Electrical Engineering Dept.
Pittsburg, Pa. 15213

Central Intelligence Agency
Attn: OCR/DD Publications
Washington, D. C. 20505

Chief of Naval Operations
OP-07
Washington, D. C. 20350 [2]

Chief of Naval Research
Department of the Navy
Washington, D. C. 20360
Attn: Code 427 [3]

Commandant
U. S. Army and General Staff College
Attn: Secretary
Fort Leavenworth, Kansas 66370

Commander
Naval Air Development and
Material Center
Johnsville, Pennsylvania 18974

Commanding General
Frankford Arsenal
Attn: SMUFA-L6000 (Dr. Sidney Ross)
Philadelphia, Pa. 19137

Commandant
U. S. Army Air Defense School
Attn: Missile Sciences Div. C and S Dept.
P. O. Box 9390
Fort Bliss, Texas 79916

Commander
U. S. Naval Air Missile Test Center
Point Magu, California 93041

Commanding General
Attn: STEWS-WS-VT
White Sands Missile Range
New Mexico 88002 [2]

Commanding General
U. S. Army Electronics Command
Fort Monmouth, N. J. 07703
Attn: AMSEL-SC
RD-D
RD-G
RD-GF
RD-MAT
XL-D
XL-E
XL-C
XL-S
HL-D
HL-CT-R
HL-CT-P
HL-CT-L
HL-CT-O
HL-CT-I
HL-CT-A
NL-D
NL-A
NL-P
NL-R
NL-S
KL-D
KL-E
KL-S
KL-T
VL-D
WL-D

Commanding General
U. S. Army Material Command
Attn: AMCRD-RS-DE-E
Washington, D. C 20315

Commanding General
U. S. Army Missile Command
Attn: Technical Library
Redstone Arsenal, Alabama 35809

Commanding Officer
Naval Avionics Facility
Indianapolis, Indiana 46241

Commanding Officer
U. S. Army Limited War Laboratory
Attn: Technical Director
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Commanding Officer
U. S. Army Materials Research Agency
Watertown Arsenal
Watertown, Massachusetts 02172

Commanding Officer
U. S. Army Security Agency
Arlington Hall
Arlington, Virginia 22212

Commanding Officer and Director
U. S. Naval Underwater Sound Lab.
Fort Trumbull
New London, Conn. 06840

Defense Documentation Center
Attn: TISIA
Cameron Station, Bldg. 5
Alexandria, Virginia 22314 [20]

Det No. 6, OAR (LODAR)
Air Force Unit Post Office
Los Angeles, Calif. 90045

Director
Advanced Research Projects Agency
Department of Defense
Washington, D. C. 20301

Director for Materials Sciences
Advanced Research Projects Agency
Department of Defense
Washington, D. C. 20301

Director
Columbia Radiation Laboratory
Columbia University
538 West 120th Street
New York, New York 10027

Director
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61803

Director
Electronics Research Laboratory
University of California
Berkeley, California 94720

Director
Electronic Sciences Laboratory
University of Southern California
Los Angeles, California 90007

Director
Microwave Laboratory
Stanford University
Stanford, California 94305

Director - Inst. for Exploratory
Research
U. S. Army Electronics Command
Attn: Mr. Robert O. Parker
Executive Secretary, JSTAC
(AMSEL-XL-D)
Fort Monmouth, N. J. 07703

Director
National Security Agency
Fort George G. Meade
Maryland 20755
Attn: James T. Tippett

Director, Naval Research Laboratory
Technical Information Officer
Washington, D. C.
Attn: Code 2000 [8]

Director
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Mass. 02139

Director
Stanford Electronics Laboratories
Stanford University
Stanford, California 94305

Commanding Officer
Naval Ordnance Laboratory
Corona, California 91720

Commanding Officer
Naval Ordnance Laboratory
White Oak, Maryland 21502 [2]

Commanding Officer
Naval Ordnance Test Station
China Lake, Calif. 93555

Commanding Officer
Naval Training Device Center
Orlando, Florida 32811

Commanding Officer
Office of Naval Research Branch Office
1030 East Green Street
Pasadena, California

Commanding Officer
Office of Naval Research Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

Commanding Officer
Office of Naval Research Branch Office
495 Summer Street
Boston, Massachusetts 02210

Commanding Officer
Office of Naval Research Branch Office
207 West 24th Street
New York, New York 10011

Commanding Officer
Office of Naval Research Branch Office
Box 39, Fleet Post Office
New York 09510 [1]

Commanding Officer
U. S. Army Electronics R & D Activity
White Sands Missile Range
New Mexico 88002

Commanding Officer
U. S. Army Engineer R & D Laboratory
Attn: STINFO Branch
Fort Belvoir, Virginia 22060

Commanding Officer
U. S. Army Research Office (Durham)
Attn: CRD-AA-IP (Richard O. Ulsh)
Box CM, Duke Station
Durham, North Carolina 27706

Commanding General
USASTRATCOM
Technical Information Center
Fort Huachuca, Arizona 85613

Commanding Officer
Harry Diamond
Attn: Dr. Bernhold Altman (AMXDO-TI)
Connecticut Ave. & Van Ness St. NW
Washington, D. C. 20438

Commanding Officer
Human Engineering Laboratories
Aberdeen Proving Ground
Maryland 21005

Commanding Officer
U. S. Army Ballistics Research Lab.
Attn: V. W. Richards
Aberdeen Proving Ground
Maryland 21005

Director, USAF Project RAND
Via: Air Force Liaison Office
The RAND Corporation
1700 Main Street
Santa Monica, Calif. 90406
Attn: Library

Director
U. S. Army Engineer Geodesy,
Intelligence and Mapping
Research and Development Agency
Fort Belvoir, Virginia 22060

Director
U. S. Naval Observatory
Washington, D. C. 20390

Director, U. S. Naval Security Group
Attn: G43
3801 Nebraska Avenue
Washington, D. C. 20390

Division of Engineering and Applied
Physics
130 Pierce Hall
Harvard University
Cambridge, Massachusetts 02138

Professor A. A. Dougal, Director
Laboratories for Electronics and
Related Science Research
University of Texas
Austin, Texas 78712

ESD (ESTI)
L. G. Hanscom Field
Bedford, Mass. 01731 [2]

European Office of Aerospace Research
Shell Building
47 Rue Cantersteen
Brussels, Belgium [2]

Colonel Robert E. Fontana
Dept. of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433

General Electric Company
Research Laboratories
Schenectady, New York 12301

Professor Nicholas George
California Institute of Technology
Pasadena, California 91109

Goddard Space Flight Center
National Aeronautics and Space Admin.
Attn: Library, Documents Section
Code 252
Green Belt, Maryland 20771

Dr. John C. Hancock, Director
Electronic Systems Research Laboratory
Purdue University
Lafayette, Indiana 47907

Dr. H. Harrison, Code RRE
Chief, Electrophysics Branch
National Aeronautics and Space Admin.
Washington, D. C. 20546

Head, Technical Division
U. S. Naval Counter Intelligence
Support Center
Fairmont Building
4420 North Fairfax Drive
Arlington, Virginia 22203

Headquarters
Defense Communications Agency
The Pentagon
Washington, D. C. 20305

Dr. L. M. Hollingsworth
AFCRL (CRN)
L. G. Hanscom Field
Bedford, Massachusetts 01731

Hunt Library
Carnegie Institute of Technology
Schenely Park
Pittsburgh, Pa. 15213

The Johns Hopkins University
Applied Physics Laboratory
8621 Georgia Avenue
Silver Spring, Maryland 20910
Attn: Boris W. Kuvshinoff
Document Librarian

Lt. Col. Robert B. Kalisch
Chief, Electronics Division
Directorate of Engineering Sciences
Air Force Office of Scientific Research
Arlington, Virginia 22209 [5]

Colonel Kee
AFFSTE
Hqs. USAF
Room ID-429, The Pentagon
Washington, D. C. 20330

Dr. S. Benedict Levin, Director
Institute for Exploratory Research
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Los Alamos Scientific Laboratory
Attn: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

Librarian
U. S. Naval Electronics Laboratory
San Diego, California 95152 [2]

Lockheed Aircraft Corp.
P. O. Box 504
Sunnyvale, California 94088

Dr. I. R. Mirman
AFSC (SCT)
Andrews Air Force Base, Maryland

Lt. Col. Bernard S. Morgan
Frank J. Seiler Research Laboratory
U. S. Air Force Academy
Colorado Springs, Colorado 80912

Dr. G. J. Murphy
The Technological Institute
Northwestern University
Evanston, Illinois 60201

Mr. Peter Murray
Air Force Avionics Laboratory
Wright-Patterson AFB, Ohio 45433

NASA Lewis Research Center
Attn: Library
21000 Brookpark Road
Cleveland, Ohio 44135

NASA Scientific & Technical
Information Facility
Attn: Acquisitions Branch (S/AK/DL)
P. O. Box 33
College Park, Maryland 20740 [2]

National Science Foundation
Attn: Dr. John R. Lehmann
Division of Engineering
1800 G Street, NW
Washington, D. C. 20550

National Security Agency
Attn: R4 - James Tippet
Office of Research
Fort George G. Meade, Maryland 20755

Naval Air Systems Command
AIR 03
Washington, D. C. 20360 [2]

Naval Electronics Systems Command
ELEX 03
Falls Church, Virginia 22046 [2]

Naval Ordnance Systems Command
ORD 32
Washington, D. C. 20360 [2]

Naval Ordnance Systems Command
SHIP 035
Washington, D. C. 20360

Naval Ship Systems Command
SHIP 031
Washington, D. C. 20360

New York University
College of Engineering
New York, New York 10019

Dr. H. V. Noble
Air Force Avionics Laboratory
Wright-Patterson AFB, Ohio 45433

Office of Deputy Director
(Research and Information Rm. 3D1037)
Department of Defense
The Pentagon
Washington, D. C. 20301

Polytechnic Institute of Brooklyn
55 Johnson Street
Brooklyn, New York 11201
Attn: Mr. Jerome Fox
Research Coordination

RAD (EMLAL-1)
Griffiss AFB, New York 13442
Attn: Documents Library

Raytheon Company
Bedford, Mass. 01730
Attn: Librarian

Lt. Col. J. L. Reeves
AFSC (SCBB)
Andrews Air Force Base, Md. 20331

Dr. A. A. Dougal
Asst. Director of Research
Office of Defense Res. and Eng.
Department of Defense
Washington, D. C. 20301

Research Plans Office
U. S. Army Research Office
3045 Columbia Pike
Arlington, Virginia 22204

Dr. H. Robl, Deputy Chief Scientist
U. S. Army Research Office (Durham)
Durham, North Carolina 27706

Emil Schafer, Head
Electronics Properties Info. Center
Hughes Aircraft Company
Culver City, California 90230

School of Engineering Sciences
Arizona State University
Tempe, Arizona 85281

SAMSO (SMSDI-STINFO)
AF Unit Post Office
Los Angeles, California 90045

SSD (SSTRT/Lt. Starbuck)
AFUPO
Los Angeles, California 90045

Superintendent
U. S. Army Military Academy
West Point, New York 10996

Colonel A. Swan
Aerospace Medical Division
AMD (AMRXI)
Brooks AFB, Texas 78235

Syracuse University
Dept. of Electrical Engineering
Syracuse, New York 13210

University of California
Santa Barbara, California 93106
Attn: Library

University of Calif. at Los Angeles
Dept. of Engineering
Los Angeles, California 90024

University of Michigan
Electrical Engineering Dept.
Ann Arbor, Michigan 48104

U. S. Army Munitions Command
Attn: Technical Information Branch
Picatinney Arsenal
Dover, New Jersey 07801

U. S. Army Research Office
Attn: Physical Sciences Division
3045 Columbia Pike
Arlington, Virginia 22204

U. S. Atomic Energy Commission
Division of Technical Information Ext.
P. O. Box 62
Oak Ridge, Tenn. 37831

Dept. of Electrical Engineering
Texas Technological College
Lubbock, Texas 79409

U. S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

Major Charles Waespy
Technical Division
Deputy for Technology
Space Systems Division, AFSC
Los Angeles, California 90045

The Walter Reed Institute of Research
Walter Reed Medical Center
Washington, D. C. 20012

AFSC (SCTR)
Andrews Air Force Base
Maryland 20331

Weapons Systems Test Division
Naval Air Test Center
Patuxtent River, Maryland 20670
Attn: Library

Weapons Systems Evaluation Group
Attn: Col. Daniel W. McElwee
Department of Defense
Washington, D. C. 20305

Yale University
Engineering Department
New Haven, Connecticut 06720

Mr. Charles F. Yost
Special Asst. to the Director of Research
NASA
Washington, D. C. 20546

Dr. Leo Young
Stanford Research Institute
Menlo Park, California 94025

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Division of Engineering and Applied Physics
Harvard University
Cambridge, Massachusetts

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
ON PATTERN CLASSIFICATION ALGORITHMS -- INTRODUCTION AND SURVEY

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)
Interim technical report

5. AUTHOR(S) (First name, middle initial, last name)
Yu-Chi Ho and Ashok K. Agrawala

6. REPORT DATE
March 1968

7a. TOTAL NO. OF PAGES
52

7b. NO. OF REFS
53

8a. CONTRACT OR GRANT NO.
N00014-67-A-0298-0006 & NASA Grant

b. PROJECT NO. NGR 22-007-068

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report No. 557

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research

13. ABSTRACT

This paper attempts to lay bare the underlying ideas used in various pattern classification algorithms reported in the literature. It is shown that these algorithms can be classified according to the type of input information required and that the techniques of estimation, decision, and optimization theory can be used to effectively derive known as well as new results.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Pattern Recognition, Survey Pattern Classification, Survey Machine Learning Artificial Intelligence Computer Learning						