NRL Report 6742

# Experiments in Discrimination and Classification

J. J. Freeman

*Microwave Antennas and Components Branch*
*Electronics Division*

December 31, 1968

NAVAL RESEARCH LABORATORY
Washington, D.C.

*This document has been approved for public release and sale; its distribution is unlimited.*

CONTENTS


| | |
|---|---|
| Abstract | ii |
| Problem Status | ii |
| Authorization | ii |
| INTRODUCTION | 1 |
| DISCRIMINATION | 1 |
| EXPERIMENTAL COMPARISON OF DISCRIMINATION TECHNIQUES | 3 |
| THE CLASSIFICATION PROBLEM | 3 |
| CONCLUSION | 9 |
| REFERENCES | 12 |

## ABSTRACT

In discrimination problems, one usually must consider many variates simultaneously. A nonparametric discrimination method, based on estimating the probability densities of two populations, is now available and appears generally applicable to discrimination problems, based on the results of extensive experiments.

The nonparametric method may also be combined with Gamba's learning without teacher or classification procedure (M. G. Kendall, "Discrimination and Classification, Multivariate Analysis," in "Multivariate Analysis," Proceedings of an International Symposium held in Dayton, Ohio, June 1965, P. R. Krishnaiah, Ed., New York: Academic Press, 1966) to obtain a new classification technique which considers all variates simultaneously.

## PROBLEM STATUS

This is an interim report; work continues on other phases of the problem.

## AUTHORIZATION

NRL Problem R06-07
Project RF 001-08-41-4552

Manuscript submitted April 24, 1968.

# EXPERIMENTS IN DISCRIMINATION AND CLASSIFICATION

## INTRODUCTION

In a recently published paper, Kendall (1) discussed new methods of attacking the problems of discrimination and classification. He illustrated his methods on data which Fisher (2) used in his first paper on discrimination. Kendall's methods essentially consider the discrimination of one variate (of a multivariate sample) at a time. Since many discrimination problems can be solved only by considering all the variates simultaneously, a nonparametric discrimination method (due to Specht (3)) is discussed which is based on a nonparametric estimate of the probability density of two populations, and this discrimination procedure is compared with Kendall's—using Fisher's data.

This discussion is essentially preliminary background to an exposition of experiments on a new algorithm for classification, or cluster analysis. This algorithm, which combines the nonparametric discrimination method with a general classification procedure suggested by Gamba (4), again considers all variates jointly. As far as is known, experimental confirmation of Gamba's classification procedure has not hitherto been reported.

## DISCRIMINATION

The discrimination problem may be defined as follows. Let $A_k (k = 1, 2, \ldots, M)$ be a given set of $p$-dimensional vectors having an unknown characteristic "A" in common, and let $B_k (k = 1, 2, \ldots, M)$ represent a given set of type "B" $p$-dimensional samples. Let $x = x_1, \ldots, x_p$ be an arbitrary vector belonging to either class A or class B. The problem is to devise an accurate procedure for assigning x to either class A or class B. If one knew the probability densities of the A and B populations, then the best procedure would be to assign x to class A if $P_A(x_1, \ldots, x_p) > P_B(x_1, \ldots, x_p)$, where $P_A(x)$ is the probability of x under the hypothesis A. However, all we are given is a finite set of samples from each class. One approach, then, is to estimate the probability density of the two $p$-dimensional populations from the finite set of samples and assign x to that population for which the estimated probability of x occurring is larger.

The mathematical problem of estimating probability densities has been studied by Leadbetter (5) and Parzen (6), among others. However, for our purposes, heuristic considerations suffice. For simplicity we consider a set of one-dimensional observations $x_i$, where $i = 1, 2, \ldots, K$. We assume that the variate $x$ is continuously distributed and has the density $p(x)$. Then

$$p(x) = \int p(\xi)\delta(x - \xi)\, d\xi = \overline{\delta(x - \xi)}, \tag{1}$$

where we have omitted the limits of integration, it being henceforth understood that the integration is to be performed over the entire region of definition of the variate, and where $\delta(x)$ is the Dirac $\delta$-function.

As a first approximation one could replace the statistical mean of $\delta(x - \xi)$ by the sample mean, thereby obtaining

$$p^*(x) = \frac{1}{K} \sum_{k=1}^{K} \delta(x - x_k), \tag{2}$$

where $p^*(x)$ is the estimate of $p(x)$. To eliminate the discontinuities introduced by the discrete nature of the averaging, we replace $p^*(x)$ by its weighted average, namely,

$$p_1^*(x) = \int p^*(\xi) g(x - \xi)\, d\xi \Big/ \int g(\xi)\, d\xi, \tag{3}$$

where $g(\xi)$, an arbitrary weighting function, is chosen to be nonnegative and symmetrical.

Now, defining $F(x - \xi)$ by

$$F(x - \xi) = g(x - \xi) \Big/ \int g(\xi)\, d\xi, \tag{4}$$

and taking the weighted mean of Eq. (2) with the normalized weighting function $F(x - \xi)$, we obtain

$$p(x) = \int p^*(\xi) F(x - \xi)\, d\xi = \frac{1}{K} \sum_{k=1}^{K} F(x - x_k) \tag{5}$$

as the form of the estimate of the probability density which we will use in this report, where $F(x) = F(-x) \geqq 0$, and $\int F(x)\, dx = 1$.

Examples of $F(x)$ are:

$$F(x) = (D/\pi)(D^2 + x^2) \tag{6a}$$

and

$$F(x) = \frac{1}{\sqrt{\pi} D} e^{-x^2/D^2} \tag{6b}$$

where $D$ is an empirical parameter.

The generalization of Eq. (5) to a bivariate set of samples is

$$p^*(x, y) = \frac{1}{K} \sum_{k=1}^{K} F_1(x - x_k) F_2(y - y_k), \tag{7}$$

where $F_1(x)$ and $F_2(x)$ may have different functional forms or different parameteric values.

Specht (3) and Fix and Hodges (7, 8) have discussed and used empirical estimates of probability densities in problems of nonparametric discrimination and pattern recognition. In particular, Specht has successfully applied this technique to the classification of electrocardiograms. The following experiments in discrimination do not embody any original ideas but are described because they are a necessary preliminary to the classification experiments, not only to determine the value of certain parameters, but also to establish the necessary background.

## EXPERIMENTAL COMPARISON OF DISCRIMINATION TECHNIQUES

Table 1 gives data on two varieties of iris (taken from Fisher's paper (2)) which Kendall (1) uses to illustrate his method of discrimination. For brevity, use $x$, $y$, $z$, and $w$ to designate the variates SL, SW, PW, and PL. One is given $K$ samples $\{x_i, y_i, z_i, w_i\}$, where $i = 1, 2, \ldots, K$ from class A, and $K$ samples $\{x'_i, y'_i, z'_i, w'_i\}$, where $i = 1, 2, \ldots, K$ from class B, and an unknown sample $\{x, y, z, w\}$ to be identified.

One computes, using Eqs. (6b) and (7),

$$P_A(x, y, z, w) = \frac{1}{\pi^2 D^4 K} \sum_{i=1}^{K} \exp\left\{-\frac{1}{D^2}\left[(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 + (w - w_i)^2\right]\right\}, \quad (8)$$

which is the probability density estimate for the occurrence of the unknown sample, subject to the hypothesis that it is of type A. Similarly, one computes

$$P_B(x, y, z, w) = \frac{1}{\pi^2 D^4 K} \sum_{i=1}^{K} \exp\left\{-\frac{1}{D^2}\left[(x - x'_i)^2 + (y - y'_i)^2 + (z - z'_i)^2 + (w - w'_i)^2\right]\right\}. \quad (9)$$

The rule of discrimination is: if $P_A(x, y, z, w) > P_B(x, y, z, w)$, allot $(x, y, z, w)$ to A. If $P_A = P_B$ make no decision.

On the other hand, one may disregard the rationale leading up to Eq. (7) and instead one may interpret

$$\frac{1}{\pi^2 D^4} \exp\left\{-\frac{1}{D^2}\left[(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 + (w - w_i)^2\right]\right\}$$

as a "closeness" function which is a measure of how close $(x, y, z, w)$ is to $(x_i, y_i, z_i, w_i)$. According to Eq. (8), $P_A(x, y, z, w)$ measures the average closeness of $(x, y, z, w)$ to class A, and the criterion for discrimination is to assign the unknown to that class which is, on the average, closer to it.

To test this closeness criterion against Kendall's criterion 40 samples from each class were chosen randomly to serve as data from which the parameters of the decision criterion for the two methods were obtained. The remaining ten samples from each class were used to test the accuracy of discrimination of the two methods. The experiment was repeated ten times, with independent random selections of the 40 samples from each class being taken each time. The results are summarized in Table 2, which indicates the slight superiority of the closeness method and, also, that the choice of the parameter in Eq. (8) is not critical.

However, we note that the closeness method is applicable in general, whereas, although it is quite effective with the present data, Kendall's method would not be useful in discriminating between two distributions having the same marginal distributions for all the variates, but with different joint distributions.

## THE CLASSIFICATION PROBLEM

Suppose we are given a set of $p$-variate samples which may originate from the same or from different populations. The problem of separating this mixed set of samples into

Table 1
Measurements on Two Varieties of Iris (in cm): SL, SW, PL, and PW

| Iris Versicolor | | | | Iris Virginica | | | |
|---|---|---|---|---|---|---|---|
| SL | SW | PL | PW | SL | SW | PL | PW |
| 7.0 | 3.2 | 4.7 | 1.4 | 6.3 | 3.3 | 6.0 | 2.5 |
| 6.4 | 3.2 | 4.5 | 1.5 | 5.8 | 2.7 | 5.1 | 1.9 |
| 6.9 | 3.1 | 4.9 | 1.5 | 7.1 | 3.0 | 5.9 | 2.1 |
| 5.5 | 2.3 | 4.0 | 1.3 | 6.3 | 2.9 | 5.6 | 1.8 |
| 6.5 | 2.8 | 4.6 | 1.5 | 6.5 | 3.0 | 5.8 | 2.2 |
| 5.7 | 2.8 | 4.5 | 1.3 | 7.6 | 3.0 | 6.6 | 2.1 |
| 6.3 | 3.3 | 4.7 | 1.6 | 4.9 | 2.5 | 4.5 | 1.7 |
| 4.9 | 2.4 | 3.3 | 1.0 | 7.3 | 2.9 | 6.3 | 1.8 |
| 6.6 | 2.9 | 4.6 | 1.3 | 6.7 | 2.5 | 5.8 | 1.8 |
| 5.2 | 2.7 | 3.9 | 1.4 | 7.2 | 3.6 | 6.1 | 2.5 |
| 5.0 | 2.0 | 3.5 | 1.0 | 6.5 | 3.2 | 5.1 | 2.0 |
| 5.9 | 3.0 | 4.2 | 1.5 | 6.4 | 2.7 | 5.3 | 1.9 |
| 6.0 | 2.2 | 4.0 | 1.0 | 6.8 | 3.0 | 5.5 | 2.1 |
| 6.1 | 2.9 | 4.7 | 1.4 | 5.7 | 2.5 | 5.0 | 2.0 |
| 5.6 | 2.9 | 3.6 | 1.3 | 5.8 | 2.8 | 5.1 | 2.4 |
| 6.7 | 3.1 | 4.4 | 1.4 | 6.4 | 3.2 | 5.3 | 2.3 |
| 5.6 | 3.0 | 4.5 | 1.5 | 6.5 | 3.0 | 5.5 | 1.8 |
| 5.8 | 2.7 | 4.1 | 1.0 | 7.7 | 3.8 | 6.7 | 2.2 |
| 6.2 | 2.2 | 4.5 | 1.5 | 7.7 | 2.6 | 6.9 | 2.3 |
| 5.6 | 2.5 | 3.9 | 1.1 | 6.0 | 2.2 | 5.0 | 1.5 |
| 5.9 | 3.2 | 4.8 | 1.8 | 6.9 | 3.2 | 5.7 | 2.3 |
| 6.1 | 2.8 | 4.0 | 1.3 | 5.6 | 2.8 | 4.9 | 2.0 |
| 6.3 | 2.5 | 4.9 | 1.5 | 7.7 | 2.8 | 6.7 | 2.0 |
| 6.1 | 2.8 | 4.7 | 1.2 | 6.3 | 2.7 | 4.9 | 1.8 |
| 6.4 | 2.9 | 4.3 | 1.3 | 6.7 | 3.3 | 5.7 | 2.1 |
| 6.6 | 3.0 | 4.4 | 1.4 | 7.2 | 3.2 | 6.0 | 1.8 |
| 6.8 | 2.8 | 4.8 | 1.4 | 6.2 | 2.8 | 4.8 | 1.8 |
| 6.7 | 3.0 | 5.0 | 1.7 | 6.1 | 3.0 | 4.9 | 1.8 |
| 6.0 | 2.9 | 4.5 | 1.5 | 6.4 | 2.8 | 5.6 | 2.1 |
| 5.7 | 2.6 | 3.5 | 1.0 | 7.2 | 3.0 | 5.8 | 1.6 |
| 5.5 | 2.4 | 3.8 | 1.1 | 7.4 | 2.8 | 6.1 | 1.9 |
| 5.5 | 2.4 | 3.7 | 1.0 | 7.9 | 3.8 | 6.4 | 2.0 |
| 5.8 | 2.7 | 3.9 | 1.2 | 6.4 | 2.8 | 5.6 | 2.2 |
| 6.0 | 2.7 | 5.1 | 1.6 | 6.3 | 2.8 | 5.1 | 1.5 |
| 5.4 | 3.0 | 4.5 | 1.5 | 6.1 | 2.6 | 5.6 | 1.4 |
| 6.0 | 3.4 | 4.5 | 1.6 | 7.7 | 3.0 | 6.1 | 2.3 |
| 6.7 | 3.1 | 4.7 | 1.5 | 6.3 | 3.4 | 5.6 | 2.4 |
| 6.3 | 2.3 | 4.4 | 1.3 | 6.4 | 3.1 | 5.5 | 1.8 |
| 5.6 | 3.0 | 4.1 | 1.3 | 6.0 | 3.0 | 4.8 | 1.8 |
| 5.5 | 2.5 | 4.0 | 1.3 | 6.9 | 3.1 | 5.4 | 2.1 |
| 5.5 | 2.6 | 4.4 | 1.2 | 6.7 | 3.1 | 5.6 | 2.4 |
| 6.1 | 3.0 | 4.6 | 1.4 | 6.9 | 3.1 | 5.1 | 2.3 |
| 5.8 | 2.6 | 4.0 | 1.2 | 5.8 | 2.7 | 5.1 | 1.9 |
| 5.0 | 2.3 | 3.3 | 1.0 | 6.8 | 3.2 | 5.9 | 2.3 |
| 5.6 | 2.7 | 4.2 | 1.3 | 6.7 | 3.3 | 5.7 | 2.5 |
| 5.7 | 3.0 | 4.2 | 1.2 | 6.7 | 3.0 | 5.2 | 2.3 |
| 5.7 | 2.9 | 4.2 | 1.3 | 6.3 | 2.5 | 5.0 | 1.9 |
| 6.2 | 2.9 | 4.3 | 1.3 | 6.5 | 3.0 | 5.2 | 2.0 |
| 5.1 | 2.5 | 3.0 | 1.1 | 6.2 | 3.4 | 5.4 | 2.3 |
| 5.7 | 2.8 | 4.1 | 1.3 | 5.9 | 3.0 | 5.1 | 1.8 |

Table 2
Comparison of the Closeness and Kendall Methods

| Closeness Method | | | | Kendall Method | | |
|---|---|---|---|---|---|---|
| D | Failures | No Decision | Successes | Failures | No Decision | Successes |
| 0.2 | 13 | 0 | 187 | 15 | 7 | 178 |
| 0.4 | 12 | 0 | 188 | | | |
| 0.6 | 11 | 0 | 189 | | | |
| 0.8 | 11 | 0 | 189 | | | |
| 1.4 | 14 | 0 | 186 | | | |

subsets all of whose elements originate from the same population is termed the "classification problem" by Kendall (1) and "self-learning," or " learning without teacher" by Gamba (4). One recalls that, in the discrimination or learning with teacher problem, one is given, say, $N_A$ samples of class A, $\{x_i\}$, where $i = 1, 2, \ldots, N_A$, $N_B$ samples of $B$, $\{x_j'\}$, where $j = 1, 2, \ldots, N_B$, and the sample $x$ to be classified. One constructs a decision rule or algorithm $F(\{x_i\}, \{x_j'\}, x)$, whose numerical value depends on the sets $A$, $B$, and the sample $x$.

If $F(\{x_i\}, \{x_j\}, x) > 0$, then one assigns $x$ to $A$. If $F(\{x_i\}, \{x_j\}, x) = 0$, one makes no decision. If $F(\{x_i\}, \{x_j'\}, x) < 0$, one assigns $x$ to $B$. For example, the nearest-neighbor rule determines which element in either of the sets $\{x_i\}$ or $\{x_j'\}$ has the smallest Euclidean distance from $x$. If that element belongs to $A$ or $B$, then $F$ is assigned, respectively, the value 1 or -1. If one is given an infinite set of samples so that, in effect, one is given the probability density for the populations $A$ and $B$, then the maximum-likelihood rule is available, namely, $F = P_A(x) - P_B(x)$.

In general, there are many discrimination rules which are more or less successful, depending upon the problem attacked. In 1961 Gamba (4) described a device which employed a novel discrimination rule to perform learning or discrimination. Gamba also described an ingenious procedure whereby the classification problem may be resolved into an iterated sequence of discrimination problems.

Gamba's procedure is to separate, at random, the given mixed set of $N$ samples into two sets of $N/2$ samples each and to designate these sets by $A_0$ and $B_0$. Treating these arbitrarily chosen sets as though they were indeed two sets of samples having different distributions, each of the $N$ samples is classified as to whether it belongs to $A_0$ or $B_0$, according to Gamba's discrimination rule. Thus, let $F(A_0, B_0, x)$ be the discrimination rule, and let $x$ be $x_1$, the first sample in $A_0$. One computes $F(A_0', B_0, x_1)$, where $A_0'$ comprises the set $A_0$ minus the sample $x_1$. If $F(A_0', B_0, x_1) > 0$, one classifies $x_1$ as class A. Similarly, one proceeds with the remaining samples of $A_0$ and of $B_0$.

After all $N$ samples are so classified, let all those just classified as class A comprise the new set $A_1$, and let the remaining samples comprise the set $B_1$. Now let $F(A_1, B_1, x)$ be the discrimination rule, and, again, letting $x$ in turn represent each of the $N$ samples, one reclassifies all the samples into two revised classes, $A_2$ and $B_2$. This procedure is repeated until, hopefully, after $n$ cycles, $A_n$ and $B_n$ are more homogeneous than $A_1$ and $B_1$.

Gamba closed his paper with the remark that experiments on this type of learning are in progress and will be reported later. Apparently, however, as previously

mentioned, these experiments were not reported, presumably because they were unsuccessful. The following experiments use the Gamba procedure with the modification that the closeness criterion is used as the discrimination rule.

In the experiments described below, based on the data in Table 1, the discrimination rule chosen was $P_A(x, y, z, w) - P_B(x, y, z, w)$, where

$$P_A = \frac{1}{N_A} \sum_j{}' \exp \left\{ - \left[ \left( \frac{x - x_j}{D} \right)^2 + \left( \frac{y - y_j}{0.75D} \right)^2 + \left( \frac{z - z_j}{D} \right)^2 + \left( \frac{w - w_j}{0.5D} \right)^2 \right] \right\} \tag{10a}$$

and

$$P_B = \frac{1}{N_B} \sum_j{}' \exp \left\{ - \left[ \left( \frac{x - x_j'}{D} \right)^2 + \left( \frac{y - y_j'}{0.75D} \right)^2 + \left( \frac{z - z_j'}{D} \right)^2 + \left( \frac{w - w_j'}{0.5D} \right)^2 \right] \right\} \tag{10b}$$

The prime in the summation of Eq. (10a) signifies that the summation is to be taken over all sample values in class A, except the sample which is being classified, if the sample being examined belongs to class A. The prime in Eq. (10b) has a similar significance. The number of different values of $j$ participating in the summation of Eqs. (10a) and (10b) are $N_A$ and $N_B$, respectively. The fact that the range of $w$ is roughly 1/2 that of $x$ or $z$ and that the range of $y$ is 3/4 that of $x$ or $z$ is reflected in the choice of the averaging coefficients in Eq. (10a) and (10b). In effect, $P_A$ and $P_B$ are proportional to the mean generalized "closeness" of the point to be classified from its deleted A and B neighborhoods, respectively.

The classification procedure consisted of the following steps: (a) of the 100 samples shown in Table 1, 50 were chosen randomly and designated as class $A_0$, and the remainder were allotted to class $B_0$, (b) for each element, $P_{A_0}$ and $P_{B_0}$ were calculated from Eqs. (10a) and (10b). Following this, each element was assigned to class $A_1$ or $B_1$, depending upon whether or not $P_{A_0} \geqq P_{B_0}$, and (c) step (b) was repeated nine times for a total of ten discriminations.

Although the results of the discrimination experiments (Table 2) indicate that the value of the weighting parameter in Eqs. (10a) and (10b) were noncritical, the range of $D$ was much more restricted in the classification experiments.

In the following experiments to determine the useful range of the parameter $D$, the initial dichotomy always consisted of set $A_0$ having 29 iris versicolor, which we designate by $29^+$, and 21 iris virginica, which we designate by $21^-$. The set $B_0$ always consisted of $29^-$ and $21^+$, namely, the remaining 50 samples. The results of the experiment are in Table 3, where, for each value of $D$, the composition of the two sets after each cycle of learning and reappraisal is enumerated. In the last experiment the initial dichotomy consisted of two perfectly homogeneous sets chosen for testing purposes.

In each learning cycle, each of the 100 samples is assigned to one of the two sets. The percentage of correct assignments which are made is a useful index of learning and is designated %S. Table 4, which gives this percentage for the successive learning cycles for various values of $D$, indicates that for $D = 0.2$, $0.3$ and $0.4$, the learning process begins oscillating after three cycles. For $D = 0.6$ the discrimination capability increases slowly for seven cycles before it reaches its final value. For $0.8 = D = 1.2$ the discrimination capability is essentially the same, reaching equilibrium after three or four cycles.

Table 3
The Changing Composition, With Learning, of Two Sets*

| Cycle Number | Composition A + | A - | B + | B - |
|---|---|---|---|---|
| $D = 0.2$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 28 | 19 | 22 | 31 |
| 2 | 32 | 21 | 18 | 29 |
| 3 | 30 | 22 | 20 | 28 |
| 4 | 33 | 22 | 17 | 28 |
| 5 | 30 | 22 | 20 | 28 |
| 6 | 33 | 22 | 17 | 28 |
| 7 | 30 | 22 | 20 | 28 |
| 8 | 33 | 22 | 17 | 28 |
| 9 | 30 | 22 | 20 | 28 |
| 10 | 33 | 22 | 17 | 28 |
| $D = 0.3$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 35 | 20 | 15 | 30 |
| 2 | 38 | 20 | 12 | 28 |
| 3 | 35 | 19 | 15 | 31 |
| 4 | 36 | 21 | 14 | 29 |
| 5 | 35 | 19 | 15 | 31 |
| 6 | 35 | 21 | 15 | 29 |
| 7 | 36 | 19 | 14 | 31 |
| 8 | 32 | 21 | 18 | 29 |
| 9 | 34 | 20 | 16 | 30 |
| 10 | 33 | 21 | 17 | 29 |
| $D = 0.4$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 40 | 18 | 10 | 32 |
| 2 | 45 | 15 | 5 | 35 |
| 3 | 44 | 11 | 6 | 39 |
| 4 | 44 | 13 | 6 | 37 |
| 5 | 44 | 11 | 6 | 39 |
| 6 | 44 | 13 | 6 | 37 |
| 7 | 44 | 11 | 6 | 39 |
| 8 | 44 | 13 | 6 | 37 |
| 9 | 44 | 11 | 6 | 39 |
| 10 | 44 | 13 | 6 | 37 |

| Cycle Number | Composition A + | A - | B + | B - |
|---|---|---|---|---|
| $D = 0.6$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 47 | 14 | 3 | 36 |
| 2 | 47 | 9 | 3 | 41 |
| 3 | 47 | 8 | 3 | 42 |
| 4 | 47 | 7 | 3 | 43 |
| 5 | 47 | 6 | 3 | 44 |
| 6 | 47 | 4 | 3 | 46 |
| 7 | 47 | 4 | 3 | 46 |
| 8 | 47 | 4 | 3 | 46 |
| 9 | 47 | 4 | 3 | 46 |
| 10 | 47 | 4 | 3 | 46 |
| $D = 0.8$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 47 | 9 | 3 | 41 |
| 2 | 48 | 3 | 2 | 47 |
| 3 | 48 | 4 | 2 | 46 |
| 4 | 48 | 4 | 2 | 46 |
| 5 | 48 | 4 | 2 | 46 |
| 6 | 48 | 4 | 2 | 46 |
| 7 | 48 | 4 | 2 | 46 |
| 8 | 48 | 4 | 2 | 46 |
| 9 | 48 | 4 | 2 | 46 |
| 10 | 48 | 4 | 2 | 46 |
| $D = 1.0$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 47 | 7 | 3 | 43 |
| 2 | 48 | 4 | 2 | 46 |
| 3 | 48 | 4 | 2 | 46 |
| 4 | 48 | 4 | 2 | 46 |
| 5 | 48 | 4 | 2 | 46 |
| 6 | 48 | 4 | 2 | 46 |
| 7 | 48 | 4 | 2 | 46 |
| 8 | 48 | 4 | 2 | 46 |
| 9 | 48 | 4 | 2 | 46 |
| 10 | 48 | 4 | 2 | 46 |

| Cycle Number | Composition A + | A - | B + | B - |
|---|---|---|---|---|
| $D = 1.2$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 48 | 9 | 2 | 41 |
| 2 | 48 | 4 | 2 | 46 |
| 3 | 48 | 4 | 2 | 46 |
| 4 | 48 | 4 | 2 | 46 |
| 5 | 48 | 4 | 2 | 46 |
| 6 | 48 | 4 | 2 | 46 |
| 7 | 48 | 4 | 2 | 46 |
| 8 | 48 | 4 | 2 | 46 |
| 9 | 48 | 4 | 2 | 46 |
| 10 | 48 | 4 | 2 | 46 |
| $D = 1.4$ | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 48 | 8 | 2 | 42 |
| 2 | 49 | 6 | 1 | 44 |
| 3 | 49 | 6 | 1 | 44 |
| 4 | 49 | 6 | 1 | 44 |
| 5 | 49 | 6 | 1 | 44 |
| 6 | 49 | 6 | 1 | 44 |
| 7 | 49 | 6 | 1 | 44 |
| 8 | 49 | 6 | 1 | 44 |
| 9 | 49 | 6 | 1 | 44 |
| 10 | 49 | 6 | 1 | 44 |
| $D = 0.3$ Initial States Perfectly Pure | | | | |
| 0 | 50 | 0 | 0 | 50 |
| 1 | 48 | 2 | 2 | 48 |
| 2 | 49 | 2 | 1 | 48 |
| 3 | 49 | 2 | 1 | 48 |
| 4 | 49 | 2 | 1 | 48 |
| 5 | 49 | 2 | 1 | 48 |
| 6 | 49 | 2 | 1 | 48 |
| 7 | 49 | 2 | 1 | 48 |
| 8 | 49 | 2 | 1 | 48 |
| 9 | 49 | 2 | 1 | 48 |
| 10 | 49 | 2 | 1 | 48 |

*Each experiment starts with the same arbitrary dichotomy of two mixed sets but employs different values of D in the discrimination criterion.

For $D = 1.4$, the percentage of successful assignments begins to drop slightly. It is surprising that the initial percentage of successful assignments (88 to 90) should be so large, while the learning models (the initial states of the two sets) are so inhomogeneous; it is also surprising that the further increase in discrimination capability is so small when, after the first learning cycle or so, the revised learning sets have become fairly pure.

Another measure of performance is the relative purity of a set, designated by $R$ and defined as the ratio of the majority to the total number of elements in the set. Table 4 gives the percentage of successful assignments %S and the relative purity $R_A$ and $R_B$ of each of the two sets, as a function of learning cycle, for different values of $D$. The same initial states of the two sets are kept, namely, an initial relative purity of 0.58. The data indicate that the average relative purity $(R_A + R_B)/2$ is approximately equal to the fractional number of successful classifications made. An indication

Table 4

The Evolution With Learning of the Percentage of Correct Discriminations and Relative Purity, for Different Values of D. The Initial Dichotomy of the Two Mixed Sets is the Same.

| D | Cycle | Cycle Number in the Classification Experiment | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.2 | %S | | 59 | 61 | 58 | 61 | | | | | | |
| 0.2 | $R_A$ | 0.58 | 0.54 | 0.63 | 0.58 | 0.59 | Rep* 3 | Rep 4 | Rep 3 | Rep 4 | Rep 3 | Rep 4 |
| 0.2 | $R_B$ | 0.58 | 0.59 | 0.62 | 0.59 | 0.62 | | | | | | |
| 0.3 | %S | | 65 | 66 | 66 | 65 | 66 | 64 | 67 | 61 | 64 | 62 |
| 0.3 | $R_A$ | 0.58 | 0.64 | 0.73 | 0.64 | 0.63 | 0.65 | 0.63 | 0.65 | 0.61 | 0.63 | 0.61 |
| 0.3 | $R_B$ | 0.58 | 0.67 | 0.70 | 0.66 | 0.66 | 0.68 | 0.66 | 0.69 | 0.62 | 0.65 | 0.63 |
| 0.4 | %S | | 72 | 80 | 83 | 81 | 83 | | | | | |
| 0.4 | $R_A$ | 0.58 | 0.69 | 0.75 | 0.80 | 0.77 | 0.80 | Rep 4 | Rep 3 | Rep 4 | Rep 3 | Rep 4 |
| 0.4 | $R_B$ | 0.58 | 0.76 | 0.88 | 0.87 | 0.86 | 0.87 | | | | | |
| 0.6 | %S | | 83 | 84 | 89 | 90 | 91 | 93 | 93 | | | |
| 0.6 | $R_A$ | 0.58 | 0.77 | 0.84 | 0.86 | 0.87 | 0.89 | 0.92 | 0.92 | Rep 7 | Rep 7 | Rep 7 |
| 0.6 | $R_B$ | 0.58 | 0.92 | 0.93 | 0.93 | 0.94 | 0.94 | 0.94 | 0.94 | | | |
| 0.8 | %S | | 88 | 95 | 94 | 94 | | | | | | |
| 0.8 | $R_A$ | 0.58 | 0.84 | 0.94 | 0.92 | 0.92 | Rep 4 | Rep 4 | Rep 4 | Rep 4 | Rep 4 | Rep 4 |
| 0.8 | $R_B$ | 0.58 | 0.93 | 0.96 | 0.96 | 0.96 | | | | | | |
| 1.0 | %S | | 90 | 94 | 94 | | | | | | | |
| 1.0 | $R_A$ | 0.58 | 0.87 | 0.87 | 0.92 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 |
| 1.0 | $R_B$ | 0.58 | 0.94 | 0.96 | 0.96 | | | | | | | |
| 1.2 | %S | | 89 | 94 | 94 | | | | | | | |
| 1.2 | $R_A$ | 0.58 | 0.84 | 0.92 | 0.92 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 |
| 1.2 | $R_B$ | 0.58 | 0.95 | 0.96 | 0.96 | | | | | | | |
| 1.4 | %S | | 90 | 93 | 93 | | | | | | | |
| 1.4 | $R_A$ | 0.58 | 0.86 | 0.89 | 0.89 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 | Rep 3 |
| 1.4 | $R_B$ | 0.58 | 0.96 | 0.98 | 0.98 | | | | | | | |
| | Initial States Perfectly Pure | | | | | | | | | | | |
| 0.3 | %S | | 96 | 97 | 97 | | | | | | | |
| 0.3 | $R_A$ | 1.0 | 0.96 | 0.98 | 0.98 | Rep | Rep | Rep | Rep | Rep | Rep | Rep |
| 0.3 | $R_B$ | 1.0 | 0.96 | 0.98 | 0.98 | | | | | | | |

*In Tables 4, 5, and 6, Rep indicates that the cycle repeats indefinitely.

of how accurately the classification would proceed if the initial dichotomy had resulted in two perfectly homogeneous sets is given by the last three lines of Table 4. For $D = 0.8$ the third and fourth cycles appear identical, as shown in Table 4. Actually, only the conclusions are identical, since the probability densities (not indicated) are different; that is, the degree of confidence is different for the two cycles. However, the fifth cycle repeated the fourth cycle precisely, and is so indicated.

After the constant value $D = 1$ was chosen, ten experiments were performed similar to the preceding ones. In these experiments the initial dichotomy, chosen by random numbers, was different in each case. Table 5 gives the composition of the two sets as the classification procedure evolves. Table 6 gives the percent of successful discriminations made in each state and the resulting relative purity of the two sets.

Experiment 7 is especially interesting. Here the initial purity was 0.52, and this increased after the first cycle to 0.58. Nevertheless, poor discrimination reduced the purity in the second cycle to 0.52, and in the following cycles the two sets were rendered increasingly homogeneous, with final relative purities of 0.96 and 0.92; however, the final majority type of elements was opposite to the initial majority type, as shown in Table 5. It is also noteworthy that experiment 10 was successfully conducted to an average final relative purity of 0.89, despite the fact that the initial learning sets had the same number of type A as of type B. Of the ten experiments performed for $D = 1$, five achieved a final successful discrimination of 94%, and the other five achieved 86%. The relative purity of the initial dichotomy of the first five was 0.568, that of the second five was 0.54, the ratios for discrimination and for purity begin roughly the same.

It seems intuitively obvious that the detailed structure of the function $F(x)$ defined in Eq. (5) is inconsequential, and that the particular form of $F(x)$ defined in Eq. (6a) would be just as effective in the classification procedure as the functional form defined by Eq. (6b). This was verified experimentally by using the same initial dichotomy for the two functional forms and obtaining essentially the same results.

## CONCLUSION

A "closeness" discrimination criterion was combined with a "learning without teacher" concept described by Gamba to successfully perform experiments in classification.

Table 5
The Evolving Composition, with Learning, of* Two Sets Initially Formed by an Arbitrary Dichotomy of Two Mixed Sets

| Cycle Number | Composition | | | |
|---|---|---|---|---|
| | A | | B | |
| | + | - | + | - |
| **Experiment 1** | | | | |
| 0 | 29 | 21 | 21 | 29 |
| 1 | 47 | 7 | 3 | 43 |
| 2 | 48 | 4 | 2 | 46 |
| 3 | 48 | 4 | 2 | 46 |
| 4 | Rep | Rep | Rep | Rep |
| **Experiment 2** | | | | |
| 0 | 28 | 22 | 22 | 28 |
| 1 | 50 | 28 | 0 | 22 |
| 2 | 50 | 20 | 0 | 30 |
| 3 | 50 | 14 | 0 | 36 |
| 4 | 50 | 14 | 0 | 36 |
| 5 | Rep | Rep | Rep | Rep |
| **Experiment 3** | | | | |
| 0 | 24 | 26 | 26 | 24 |
| 1 | 27 | 36 | 23 | 14 |
| 2 | 24 | 44 | 26 | 6 |
| 3 | 18 | 47 | 32 | 3 |
| 4 | 13 | 48 | 37 | 2 |
| 5 | 7 | 48 | 43 | 2 |
| 6 | 3 | 47 | 47 | 3 |
| 7 | 2 | 46 | 48 | 4 |
| 8 | 2 | 46 | 48 | 4 |
| 9 | Rep | Rep | Rep | Rep |
| **Experiment 4** | | | | |
| 0 | 24 | 26 | 26 | 24 |
| 1 | 5 | 24 | 45 | 26 |
| 2 | 4 | 32 | 46 | 18 |
| 3 | 0 | 36 | 50 | 14 |
| 4 | 0 | 36 | 50 | 14 |
| 5 | Rep | Rep | Rep | Rep |
| **Experiment 5** | | | | |
| 0 | 20 | 30 | 30 | 20 |
| 1 | 8 | 31 | 42 | 19 |
| 2 | 4 | 34 | 46 | 16 |
| 3 | 0 | 36 | 50 | 14 |
| 4 | 0 | 36 | 50 | 14 |
| 5 | Rep | Rep | Rep | Rep |
| **Experiment 6** | | | | |
| 0 | 24 | 26 | 26 | 24 |
| 1 | 1 | 33 | 49 | 17 |
| 2 | 0 | 36 | 50 | 14 |
| 3 | 0 | 36 | 50 | 14 |
| 4 | Rep | Rep | Rep | Rep |
| **Experiment 7** | | | | |
| 0 | 26 | 24 | 24 | 26 |
| 1 | 29 | 21 | 21 | 29 |
| 2 | 25 | 23 | 25 | 27 |
| 3 | 23 | 33 | 27 | 17 |
| 4 | 20 | 43 | 30 | 7 |
| 5 | 15 | 46 | 35 | 4 |
| 6 | 9 | 48 | 41 | 2 |
| 7 | 4 | 48 | 46 | 2 |
| 8 | 2 | 47 | 48 | 3 |
| 9 | 2 | 46 | 48 | 4 |
| 10 | 2 | 46 | 48 | 4 |
| **Experiment 8** | | | | |
| 0 | 22 | 28 | 28 | 22 |
| 1 | 8 | 35 | 42 | 15 |
| 2 | 3 | 47 | 47 | 3 |
| 3 | 2 | 47 | 48 | 3 |
| 4 | 2 | 46 | 48 | 4 |
| 5 | 2 | 46 | 48 | 4 |
| 6 | Rep | Rep | Rep | Rep |
| **Experiment 9** | | | | |
| 0 | 17 | 33 | 33 | 17 |
| 1 | 5 | 47 | 45 | 3 |
| 2 | 2 | 46 | 48 | 4 |
| 3 | 2 | 46 | 48 | 4 |
| 4 | Rep | Rep | Rep | Rep |
| **Experiment 10** | | | | |
| 0 | 25 | 25 | 25 | 25 |
| 1 | 12 | 23 | 38 | 27 |
| 2 | 11 | 27 | 39 | 23 |
| 3 | 7 | 29 | 43 | 21 |
| 4 | 4 | 32 | 46 | 18 |
| 5 | 0 | 36 | 50 | 14 |
| 6 | 0 | 36 | 50 | 14 |
| 7 | Rep | Rep | Rep | Rep |

*For each experiment $D = 1$, but the initial dichotomy is different.

Table 6
The Evaluation with Learning of the Percentage of Correct Discrimination and Relative Purity of Two Sets Formed Initially by an Arbitrary Dichotomy of Two Mixed Sets. For Each Experiment $D = 1$, but the Initial Dichotomy is Different.

| Experiment Number | Cycle | Cycle Number in the Classification Experiment | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | %S | | 90 | 94 | 94 | Rep | | | | | | |
| | $R_A$ | 0.58 | 0.87 | 0.87 | 0.92 | Rep | | | | | | |
| | $R_B$ | 0.58 | 0.94 | 0.96 | 0.96 | Rep | | | | | | |
| 2 | %S | | 72 | 80 | 86 | 86 | Rep | | | | | |
| | $R_A$ | 0.56 | 0.64 | 0.71 | 0.78 | 0.78 | Rep | | | | | |
| | $R_B$ | 0.56 | 1.0 | 1.0 | 1.0 | 1.0 | Rep | | | | | |
| 3 | %S | | 59 | 70 | 79 | 85 | 91 | 94 | 94 | 94 | Rep | |
| | $R_A$ | 0.52 | 0.57 | 0.65 | 0.72 | 0.79 | 0.87 | 0.94 | 0.96 | 0.96 | Rep | |
| | $R_B$ | 0.52 | 0.62 | 0.81 | 0.91 | 0.95 | 0.96 | 0.94 | 0.92 | 0.92 | Rep | |
| 4 | %S | | 69 | 78 | 86 | Rep | | | | | | |
| | $R_A$ | 0.52 | 0.83 | 0.89 | 1.0 | Rep | | | | | | |
| | $R_B$ | 0.52 | 0.63 | 0.72 | 0.78 | Rep | | | | | | |
| 5 | %S | | 73 | 80 | 86 | 86 | Rep | | | | | |
| | $R_A$ | 0.60 | 0.80 | 0.90 | 1.0 | 1.0 | Rep | | | | | |
| | $R_B$ | 0.60 | 0.69 | 0.74 | 0.78 | 0.78 | Rep | | | | | |
| 6 | %S | | 82 | 86 | 86 | Rep | | | | | | |
| | $R_A$ | 0.52 | 0.97 | 1.0 | 1.0 | Rep | | | | | | |
| | $R_B$ | 0.52 | 0.74 | 0.78 | 0.78 | Rep | | | | | | |
| 7 | %S | | 58 | 52 | 40 | 27 | 19 | 11 | 6 | 5 | 6 | 6 |
| | $R_A$ | 0.52 | 0.58 | 0.52 | 0.41 | 0.32 | 0.25 | 0.16 | 0.08 | 0.04 | 0.04 | 0.04 |
| | $R_B$ | 0.52 | 0.58 | 0.52 | 0.39 | 0.19 | 0.10 | 0.05 | 0.04 | 0.06 | 0.08 | 0.08 |
| 8 | %S | | 77 | 94 | 95 | 94 | 94 | Rep | | | | |
| | $R_A$ | 0.56 | 0.81 | 0.94 | 0.96 | 0.96 | 0.96 | Rep | | | | |
| | $R_B$ | 0.56 | 0.74 | 0.94 | 0.94 | 0.92 | 0.92 | Rep | | | | |
| 9 | %S | | 92 | 94 | 94 | Rep | | | | | | |
| | $R_A$ | 0.66 | 0.90 | 0.96 | 0.96 | Rep | | | | | | |
| | $R_B$ | 0.60 | 0.94 | 0.92 | 0.92 | Rep | | | | | | |
| 10 | %S | | 61 | 66 | 72 | 78 | 86 | 86 | Rep | | | |
| | $R_A$ | 0.50 | 0.66 | 0.71 | 0.81 | 0.89 | 1.0 | 1.0 | Rep | | | |
| | $R_B$ | 0.50 | 0.59 | 0.63 | 0.67 | 0.72 | 0.78 | 0.78 | Rep | | | |

## REFERENCES


1. Kendall, M. G., "Discrimination and Classification, Multivariate Analysis," in "Multivariate Analysis," Proceedings of an International Symposium held in Dayton, Ohio, June 1965, P. R. Krishnaiah, ed., New York: Academic Press, 1966

2. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Ann. Eugenics 7:179-188 (1936)

3. Specht, D. F., "Generation of Polynomial Discriminant Functions for Pattern Recognition," Technical Report 6764-5, Stanford Electronics Laboratories, Stanford University, May 1966

4. Gamba, A., Palmieri, G., and Sanna, R., "Self Learning in PAPA," Suppl. Nuovo Cimento 20:146-147 (1961)

5. Leadbetter, M. R., "On the Nonparametric Estimation of Probability Densities," Technical Report 11, Research Triangle Institute, May 24, 1963

6. Parzen, E., "On Estimation of a Probability Density Function and Mode," Ann. Math. Stat. 33:1065-1076 (1962)

7. Fix, E., and Hodges, J. L., Jr., "Nonparametric Discrimination: I Consistency Properties," Report 4, Project 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, Feb. 1951

8. Fix, E., and Hodges, J. L., Jr., "Nonparametric Discrimination: Small Sample Performance," Report 11, Project 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, Aug. 1952

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)<br>Naval Research Laboratory<br>Washington, D.C. 20390 | 2a. REPORT SECURITY CLASSIFICATION<br>Unclassified<br>2b. GROUP |

3. REPORT TITLE

EXPERIMENTS IN DISCRIMINATION AND CLASSIFICATION

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

An interim report; work on other phases of the problem continues.

5. AUTHOR(S) (First name, middle initial, last name)

J. J. Freeman

| | | |
|---|---|---|
| 6. REPORT DATE<br>December 31, 1968 | 7a. TOTAL NO. OF PAGES<br>16 | 7b. NO. OF REFS<br>8 |
| 8a. CONTRACT OR GRANT NO.<br>NRL Problem R06-07<br>b. PROJECT NO.<br>RF 001-08-41-4552<br>c.<br>d. | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>NRL Report 6742<br>9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) | |

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

| | |
|---|---|
| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY<br>Department of the Navy (Office of Naval Research), Washington, D.C. 20360 |

13. ABSTRACT


In discrimination problems, one usually must consider many variates simultaneously. A nonparametric discrimination method, based on estimating the probability densities of two populations is now available and appears generally applicable to discrimination problems, based on the results of extensive experiments.

The nonparametric method may also be combined with Gamba's learning without teacher or classification procedure (M.G. Kendall, "Discrimination and Classification, Multivariate Analysis," in "Multivariate Analysis," Proceedings of an International Symposium held in Dayton, Ohio, June 1965, P.R. Krishnaiah, Ed., New York:Academic Press, 1966) to obtain a new classification technique which considers all variates simultaneously.


DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801



Security Classification

Security Classification

| 14. KEY WORDS | LINK A ROLE | LINK A WT | LINK B ROLE | LINK B WT | LINK C ROLE | LINK C WT |
|---|---|---|---|---|---|---|
| Discrimination | | | | | | |
| Classification | | | | | | |
| Machine learning | | | | | | |
| Pattern, recognition | | | | | | |
| Artificial intelligence | | | | | | |