AD-A051 549 NAVAL UNDERSEA RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 17/1
AN APPROACH TO TARGET CLASSIFICATION BY COMPUTER IN ADVANCED AC--ETC(U)
JUN 71 J A ROESE , G A BUTLER

UNCLASSIFIED NUC-TN-542 NL

| OF |
AD
A051549

END
DATE
FILMED
4-78
DDC

MOST Project 4 - 3

(14) NUC-TN-542

(1)

(9) Technical note Jun 67 - Dec 68

# NAVAL UNDERSEA RESEARCH AND DEVELOPMENT CENTER

(6) AN APPROACH TO TARGET CLASSIFICATION BY COMPUTER IN ADVANCED ACTIVE SONAR SYSTEMS

Good.

by

(10) J. A. Roese and G. A. Butler



(11) June 1971

(12) 52p.

San Diego, California

[redacted]

## FOREWORD

This technical note suggests methods for incorporation of automatic and semi-automatic target classification techniques into the design of advanced active sonar systems. The discussion results in a proposed design for a classification subsystem that is computationally fast, adaptive, and provides operationally meaningful information to sonar, fire control, and command and control personnel.

The work represented herein was done from June 1967 through December 1968 under direction of NUC personnel now associated with NUC Code 603, Simulation, Analysis and Applications Division. It was sponsored by NAVSHIPSYSCOM (Code 00V2) through the Conformal/Planar Array Program Project Office and the New Submarine Sonar/Fire Control System Project Office. Technical assistance was obtained from Computer Applications, Incorporated (CAI), under Contract N123(953)57400A.

Only a portion of the total effort expended is reported in this technical note. An alternative nonstatistical approach was also pursued. The results of this latter work are documented in NUC Technical Note 543.[1]

The guidance and technical contributions of R. P. Schindler, now of the Naval Electronics Laboratory Center, are gratefully acknowledged, as well as the programming support provided by Mrs. J. Sentovic and R. T. Napier, also of that Center, and by M. Einhorn of Computer Sciences Corporation.



---

[1] Superscript numbers denote references at end of report, preceding the appendices.

CONTENTS

## TABLES



## ILLUSTRATIONS

# AN APPROACH TO TARGET CLASSIFICATION BY COMPUTER IN ADVANCED ACTIVE SONAR SYSTEMS

## I. INTRODUCTION

Traditionally, and to a large extent today, the responsibility for making a classification decision on a detected target rests with the sonar operator. It is his task to review all target-related information received through the various inputs available to him. From these sources he must extract only the information he considers pertinent and then correlate this information before arriving at a classification decision.

The need is great for the generation of automated techniques to assist the sonar operator in selecting, extracting and correlating pertinent data. This need will become more acute in the complex advanced sonar systems which are currently under development.

The present investigation included four kinds of overlapping activities: (1) a search of current literature for applicable techniques, (2) development of new analytical techniques to augment existing ones, (3) generation of computer programs to implement and evaluate these techniques, and (4) development of a design for a complete semi-automatic classification subsystem.

The general classification problem is formidable, and despite the fact that the approach in this investigation was restricted to methods realizable in a real-time system, the results are necessarily partial and the recommendations tentative. Nevertheless, essential ground has been covered, and it is believed

the observations made here will remain valid until the next major theoretical advances in classification techniques are achieved. Also, many of the topics discussed in this note may be of interest to those concerned with applications of general computerized learning and classification methods to areas beyond the domain of advanced digital sonar systems.

## II. CLASSIFICATION IN ADVANCED ACTIVE SONAR SYSTEMS

A. The Classification Subsystem. Advanced submarine and surface ship sonar systems now under Navy development are being designed to incorporate computerized data processing complexes (DPC). The DPC will digitally correlate the inputs from the active and passive sonar sensors and combine them with external inputs such as environmental data, own-ship's status, and intelligence information. High-speed digital processing of this information will allow the system to carry out, with varying degrees of operator assistance, the functions of target detection, tracking, classification, and threat evaluation, and will automatically provide fire control solutions and weapon settings. In addition, the DPC will generate graphic and digital display formats for sonar, fire control, and command and control personnel.

This technical note deals with the portion of the DPC that performs the target classification function utilizing the active sonar returns. It is considered a subsystem of the DPC, and is subject to normal system constraints. For instance, the classification subsystem must be real-time in the sense that it must rapidly process its external inputs and output results before data from the next ping arrive. Since a digital computer is the heart of the DPC, the classification computer programs must be designed to respond first to inputs of high priority to the exclusion or deferment of lower priority functions. Time-consuming operations and those not requiring fast response cannot be permitted to interfere, and are accordingly relegated to "background" processing. In addition to the design considerations peculiar to a real-time system, this subsystem shares the constraint of any complex computational system of programs - that a subprogram may not consume operating time or core memory space which is out of proportion to its importance.

B. Probabilities of Class Membership. The computational power available in the DPC permits the application of statistical decision theory to the problem of either automatic or semi-automatic (operator-assisted) target classification. Desirable simplifications can be made in the statistical theory if we ignore a priori probabilities and cost functions. This can be done because there is no risk involved in ranking targets by probability scalars, as opposed to deciding that they belong to class $\Theta_1$, $\Theta_2$, . . $\Theta_n$. The classification subsystem outputs are in the form of likelihood ratios or probabilities and associated confidence levels. The likelihood ratios can be expressed as

$$\ell_K = \frac{P(X|\Theta_k)}{1-P(X|\Theta_k)} \qquad (1)$$

where $P(X|\Theta_k)$ is the probability of event X occurring given that it is a member of class $\Theta_k$. To convey information to an operator in a more meaningful fashion, these probabilities can be expressed directly:

$$P(\Theta_k|X) = \frac{P(X|\Theta_k)P(\Theta_k)}{\sum_{k=i}^{all} P(X|\Theta_k)} \qquad (2)$$

where X is the unknown observation and $P(X|\Theta_k)$ is the conditional probability density for the $k^{th}$ class. Restated in these terms, the classification problem becomes one of computing $P(\Theta_k|X)$ from the estimated conditional probability densities $P(X|\Theta_k)$. To obtain acceptable estimates of these probability densities, we should observe as many samples from each target class as possible. This is extremely difficult, especially in the case of obtaining samples of actual hostile vessels and weapons.

It is clear, however, that the computer must have some estimate of each of these probability densities to make a classification. These estimates can and should be improved as more observations are made. The initial densities used will be based on simulated data, the generation of which is a difficult problem in itself. Ideally, the influence of the simulated data will gradually diminish as real samples become available, but there is little assurance that real samples would occur frequently enough or over a sufficiently representative range to obviate the simulated data completely.

C. Learning. The procedure which forms a representative conditional probability density function $P(X|\Theta_k)$ for a class of objects from past observations of samples is referred to as "learning". The most thoroughly analyzed and tractable application of statistical learning is for multivariate normal distributions (MND). Learning, for the MND, is the computation of the sample covariance matrix, U, and the sample mean, $\bar{X}$, as estimates of the parameters of the sample's parent distribution $(\Sigma,\mu)$. This exemplifies "parametric learning" where the form of the underlying distribution is assumed and its characteristic parameters are estimated with diminishing error as the number of samples increases.

Conversely, non-parametric or distribution free learning assumes nothing about the form of the distribution, but responds directly to the samples as does a histogram. The single outstanding non-parametric multivariate method existent is the Polynomial Discriminant Method (PDM).[2] The relative merits of the parametric and non-parametric methods will be discussed in Section III.

The major difficulty in applying the learning procedures to sonar is that of obtaining usable quantities of representative data over the range of circumstances in which the target types of interest occur. For submarine targets, for example, these data include pertinent combinations of range, speed, depth, and aspect angle for surface, bottom bounce, and convergence zone modes of transmission.

D. Adaptation. Adaptation is interpreted as the process of modifying the originally learned probability density functions in response to the reception of additional samples of known target classes. These samples have previously been classified by an external source. If the new samples are weighted the same as samples originally learned, the final result will be the same as if all the samples had arrived at the same time. Such an elementary scheme is appropriate only to data that are time-invariant, and is a special case of adaptation. The importance of adaptation is apparent from the following comments on the characteristics of certain target types:

1. Submarines will provide comparatively few samples. It would therefore be necessary to remain sensitive to new samples while retaining all of the older learned information. Also, the real-time between observations of submarine targets ordinarily would be great, suggesting that time-weighting should be discounted for submarine samples.

2. Noise, in contrast to submarines, would provide a far more continual supply of representative target samples which might vary gradually with time and location. Here, a straightforward time-weighting scheme would cause the influence of older samples to fade and be supplanted by the newer information. The rate at which the weight would diminish with age would be controlled by a parameter chosen empirically. The relative stability of most classes of targets would permit adaptation computations to take place on an infrequent background basis.

3. Other target sources such as sea mounts, schools of fish, whales, etc., fall between submarines and noise in the rate at which samples will be

available, and it is another problem how to merge these sources for learning.

The point here is that the method of adaptation has to be consistent with our model of the processes we are studying; i. e., are they stationary, highly unpredictable, etc.? From the point of view of simplicity, we would like the adaptation algorithm to be in parametric form and not involve elaborate computations.

E. Classification. The act of making a target classification decision is too crucial and the cost of a false alarm too high to permit this function to be completely automatic. The operator with the responsibility of making a target classification decision will have to act on his own judgment, based on the automatic estimates of the target class probabilities and tempered by all available active and passive tracking data.

For each ping's worth of new data, the classification subsystem will automatically assign class probabilities, $P(\Theta_k|X)$, for newly detected targets and for those which are currently being actively or passively tracked. Ideally, the classification subsystem would unerringly label each target with the correct $\Theta_k$. Realistically, the best that can be done is to estimate the likelihood that the unknown target is from $\Theta_k$ and also estimate a confidence level for the likelihood ratio. These variables may be used conveniently to rank the relative importance of each target as suggested by Table 1.

TABLE 1. TARGET RELATIVE IMPORTANCE AS A FUNCTION OF HOSTILE TARGET LIKELIHOOD RATIO AND CONFIDENCE IN LIKELIHOOD RATIO

| Hostile Target Likelihood Ratio | Confidence in Likelihood Ratio | Target Relative Importance |
|---|---|---|
| Low | High | Least important |
| Low | Low | Important |
| High | Low | Important |
| High | High | Most Important |

Appendix A discusses a procedure which could be used to evaluate relative target importance as suggested by Table 1.

Classification, in this context, involves the evaluation of all the $P(X|\Theta_k)$'s for some selected number of targets each ping, and then, from Equation 2, computing the desired scalar class membership probabilities, $P(\Theta_k|X)$, for these targets. The classification subsystem, therefore, serves primarily to give a ranking, in terms of a scalar quantity, $P(\Theta_k|X)$ which may be used to assist the operator in making his classification decision or may be thresholded to provide an automatic alarm. The latter approaches a totally automatic classification subsystem.

The classification function exemplifies a foreground, priority function of a real-time data processing system due to the frequency of its operation and to the fact that the sonar platform may well be in a race with a hostile computer to establish a fire control solution.

## III. TECHNICAL DISCUSSION

A. Vector Representation, the Measurement Space. For each ping's worth of available information, target observations are made which consist of extracting d ordered measurements. These measurements can be represented by d ordered real numbers $(x_1, x_2, \ldots, x_i, \ldots, x_d)$ corresponding to d-dimensional vectors, X, or points in Euclidean d-space which we will refer to as the "measurement" space. This representation is necessary for a mathematical treatment and is merely an extension of the familiar Cartesian Coordinates.

Suppose that m active sonar echoes from a target are observed and three measurements are taken from each echo: Range (R), amplitude (A), and doppler (D). Table 2 illustrates both a descriptive notation and the generalized notation which will be used throughout the remainder of this note.

TABLE 2. VECTOR NOTATION

Descriptive Notation

| OBSERVATION | R | A | D |
|---|---|---|---|
| | Dimension | | |
| | $1^{st}$ | $2^{nd}$ | $d^{th}$ |
| $1^{st}$ | $R_1$ | $A_1$ | $D_1$ |
| $2^{nd}$ | $R_2$ | $A_2$ | $D_2$ |
| . . . | . . . | . . . | . . . |
| $j^{th}$ | $R_j$ | $A_j$ | $D_j$ |
| . . . | . . . | . . . | . . . |
| $m^{th}$ | $R_m$ | $A_m$ | $D_m$ |

Generalized Notation

| j | i | | |
|---|---|---|---|
| | 1 | 2 | 3(=d) |
| 1 | $x_{11}^k$ | $x_{12}^k$ | $x_{13}^k$ |
| 2 | $x_{21}^k$ | $x_{22}^k$ | $x_{23}^k$ |
| . . . | . . . | . . . | . . . |
| j | $x_{j1}^k$ | $x_{j2}^k$ | $x_{j3}^k$ |
| . . . | . . . | . . . | . . . |
| m | $x_{m1}^k$ | $x_{m2}^k$ | $x_{md}^k$ |

R = range, A = amplitude, D = doppler

Each set of d-ordered measurements from a target source is stored as a row vector in an m x d matrix. Each row vector is denoted by $X_j^k = (x_{j1}^k, x_{j2}^k, \ldots, x_{ji}^k, \ldots, x_{jd}^k)$ where the superscript indicates that the vector is known to be from the $k^{\underline{th}}$ class. The absence of a superscript means that the source is a target of unknown class. Sets of known vectors are denoted by brackets; e. g., a set of $M_n$ known noise vectors is $[X_j^n]M_n$.

B. Discriminant Functions. The possibility of using multiple linear surfaces (discriminant functions) to compartmentalize the measurement space was considered but found to be an unsuitable approach for the following reasons:

1. Specifically separating surfaces are sensitive only to the available known samples and are susceptible to error when used to classify unknowns.

2. The problem does not call for classification decisions, but for relative estimates that an unknown target is a submarine, torpedo, etc.

3. The direct use of separating surfaces does not provide a plausible mechanism for estimating the desired likelihood ratios and confidence levels. Of course, artificial measures of the likelihood ratios and confidence levels could always be defined in terms of distances from surfaces in an arbitrary, artificial fashion.

4. Finally, surfaces determined from a statistical foundation can be made to classify at least as well as those from class-separating algorithms by comparing the likelihood ratio with a constant (C). For example (and ignoring considerations of costs and a priori probabilities), decide:

$$X \varepsilon \Theta_1 \text{ if } \frac{P(X|\Theta_1)}{P(X|\Theta_2)} > C \tag{3}$$

Or from equation 2 decide:

$$X \varepsilon \Theta_1 \text{ if } P(X|\Theta_1) > \frac{C}{(C+1)} \tag{4}$$

Since the functions $P(X|\Theta_1)$ and $P(X|\Theta_2)$ are unrestricted, the resulting decision regions can become very complex. It should be pointed out that overly complex surface structures are not only hard to use, but may well give results inferior to those obtained with simpler surfaces.

C. Clusters and the Shape of Sample Distributions. In the statistical approach selected for this analysis, known vectors can be considered as samples from unknown parent distributions. These samples may be visualized as comprising clusters of points in the familiar three-dimensional space of our experience. Thus, sample populations from a given source class may be described in general qualitative terms such as sparse, dense, locally dense, ellipsoidally symmetric, homogeneous, etc. More complicated cluster configurations are more difficult to describe verbally or in terms of mathematical parameters. The problem of efficient description of an entire sample set can be greatly alleviated if the data are divided into easily describable subsets. Such attempts are discussed later in this section.

D. Sample Representation. If a large number of sample observations have been taken on a target, it would not be practical to store all the samples explicitly for later reference. One method of alleviating this problem would be to decimate the data, but this could lead to an undesirable loss of information in cases where the residual samples are not representative of the discarded samples. Also, the data that can be discarded without undue compromise depends upon the quality and amount of the data. Therefore, the motivation to reduce the requirement for reference data must be subordinate to the problem of faithfully representing the "shape" of any clusters contained in the data. In this note all numerical attempts at data representation are be referred to as "cluster analysis".

The literature on cluster analysis is extensive, but very little is applicable to cluster analysis in sonar classification. The best computer-oriented cluster finding technique uncovered in the literature seems to be ISODATA.[3] Successful experiments have been carried out at this Center with the ISODATA technique. However, existing cluster analysis methods appeared to have too many inherent risks and deficiencies. For example, methods which introduce the data sequentially are sensitive to the order in which the data are introduced. Also, the picture can change precipitously when certain parameters are marginal; e. g., the parameter that governs whether two or more clusters should be combined into one cluster. The validity of some other methods depends upon the shape of clusters. For instance, the histogram method of G. Sebestyen favors ellipsoidal shapes.[4]

These shortcomings were deemed unacceptable because of our overall

objective of recommending techniques which would perform reasonably well over a broad range of cluster configurations. Accordingly, a method which would regard all data simultaneously and fix the number of clusters as unequivocably as possible was sought. A vector field approach of considerable promise was developed and its feasibility was verified graphically by programs using the CALCOMP plotter and a computer-driven display.[5]

The evaluation of the efficacy of a cluster-seeking method should not be merely subjective. A step toward more objective comparisons has already been made in the generation of a few "standard" sets of data. We feel that these clearly useful attempts are nevertheless incapable of meeting an essential objection: a truly general cluster analysis technique must demonstrate itself over the applicable range of sample sizes, dimensions, and, most important of all, data configurations. We approached this problem by using Monte Carlo techniques where the data sets were generated randomly. Because this implied a lack of complete control over the test data, an automatic and objective way of measuring cluster finding performance was called for. Appendix B contains suggestions for objectively measuring this performance.

It should be emphasized that a rigorous test of a cluster analysis technique is necessary before it can be considered ready for operational use. The test data should be designed to produce significant problems in contrast to the well-separated clusters often used to illustrate techniques. As a step toward the generation of test data, a program was written which makes use of randomly generated covariance matrices. This program implements suggestions

of T. P. Norris (NUC) and G.A. Butler (CAI) and carries out the following steps:

1. Randomly selects d real vectors of d components each. These are a basis if they are chosen to be linearly independent.

2. Does a Gram-Schmidt orthogonalization process using those d vectors. The resulting orthonormal set is then arrayed in the d x d matrix, W.

3. A scalar matrix, S with diagonal elements $\lambda_1, \lambda_2 \ldots \lambda_d$, all greater than zero, is chosen randomly.

4. The desired covariance matrix is then computed by the similarity transformation

$$U = W^t SW \tag{5}$$

where U is a positive definite matrix with a determinant equal to $\prod_{i=1}^{d} \lambda_i$. If U corresponds to a d-variate normal distribution, we could say the $\lambda$'s determine its shape and spread, and W determines its orientation. The idea is that test clusters of samples generated by the distribution would tend to have the same shape and orientation. Another program, TDATA,* generates a specified number of samples from U with an assumed mean of zero. We have the ability to produce random cluster systems by generating the union of a number of more or less ellipsoidal clusters whose means would be chosen to ensure overlap. The overlap in turn gives rise to larger clusters of more complex shapes which cluster analysis techniques would have to resolve into their simpler components.

---

*TDTA was borrowed from L. Traister of CAI, and modified by R. Napier and J. Roese of NUC.

E. Measurement Selection. We have not attempted to resolve the difficult question of how many measurements should be taken on a detected target. It is true that the amount of classification information does not decrease when new measurements are added and are not found to be helpful. However, if it cannot be shown that each new measurement is independent of all the others, it is a costly effort to find out just how much information it is contributing.[6] Aside from the theoretical arugument against too many measurements, the fact is inescapable that an increase in dimensionality will mean non-linear increases in oomputer storage and processing time. At this point we can only suggest that measurements be chosen which we know to be most relevant to the physics of the situation. This suggests that each mode of active sonar transmission (surface duct, bottom bounce, convergence zone) should have its own set of measurements.

F. Statistical Learning. Learning, in the context of this problem, means the estimation of all of the conditional probability density functions $P(X|\Theta_k)$ of the $\Theta_k$ target classes. It will have to be assumed that the sample data from which the system learns is representative; that is, that the relative number of samples in a unit volume of the measurement space is suggestive of the probability density there. As was indicated earlier, it is extremely unlikely that there will ever be enough real samples to meet this requirement; therefore, the use of simulated data from elaborate models is almost inevitable. By breaking up the available samples into clusters of more or less convex shape, it is assumed that each cluster consists of samples from a local d-variate normal distribution. Learning is then reduced to computing the sample mean vector, $\bar{X}^k$, and sample covariance matrix, $U^k$, for each cluster. By definition these computations are as follows:

$$\bar{x}_i^k = \frac{\sum_{j=1}^{M_k} x_j^k}{M_k} \tag{6}$$

$$\bar{X}^k = (\bar{x}_1^k, \bar{x}_2^k, \ldots, \bar{x}_d^k) \tag{7}$$

$$u_{sr}^k = u_{rs}^k = \sum_{j=1}^{M_k} ( x_{jr}^k - \bar{x}_r^k ) ( x_{js}^k - x_s^k ), \; M_k \geq d \tag{8}$$

$$U^k = (u_{rs}^k) \tag{9}$$

For the purpose of efficient computation, the means of the measurements would be computed first; then the elements of the sample covariance matrix can be computed quickly using Equation 10:

$$u_{rs}^{k} = \frac{\sum_{j=1}^{M_k} x_{jr}^{k} x_{js}^{k}}{M_k - \bar{x}_r^{k} \bar{x}_s^{k}} \tag{10}$$

It should be noted that the simple computations of Equations (6) and (10) are all that is needed to estimate the parameters of the d-variate normal distribution which has the form of Equation (11):

$$\hat{P}(x|\theta_k) = (2\pi)^{-d/2} |U^k|^{-1/2} \exp[-1/2(X - \bar{X}^k)\ U^{k^{-1}} (X - \bar{X}^k)^T] \tag{11}$$

The circumflex reminds us that Equation (11) is an approximation to some underlying probability density whose form we have not literally assumed, but which we hope is not too unlike the normal case.

The symmetrical nature of the quadratic form in the exponent of Equation (11) permits a computational shortcut with significant savings as shown below.

Let

$$Q^k = (X - \bar{X}^k)\ U^{k^{-1}}\ (X - \bar{X}^k)^T \tag{12}$$

$$= 2\sum_{r=1}^{d-1} \sum_{s=r+1}^{d} u_{rs}(x_r - x_r^{-k})\ (x_s - x_s^{-k})$$

$$+ \sum_{r=1}^{d} u_{rr}(x_r - \bar{x}_r^{k})^2 \tag{13}$$

The derivation of the multivariate normal distribution and the optimality of the sample mean and sample covariance matrix estimators are thoroughly developed in the literature.[7,8]

The computation of the likelihood ratio also has desirable simplicity. If the likelihood ratio, $\ell$, is concerned with class h versus class k, then

$$\ell = \frac{P(x|\Theta_h)}{P(x|\Theta_k)} = \frac{|U^k|^{1/2}\exp(-1/2Q^h)}{|U^h|^{1/2}\exp(-1/2Q^k)} \tag{14}$$

It is more convenient to deal instead with the natural logarithm of $\ell$ which is, of course, monotonic with respect to $\ell$.

$$\log_e \ell = 1/2(\log_e|U^k| - Q^h - \log_e|U^h| + Q^k) \tag{15}$$

$$\log_e \ell \propto (\log_e|U^k| - \log_e|U^h| + Q^k - Q^h) \tag{16}$$

The last equation shows that the evaluation of quadratic forms is all that is required to determine $\ell$, since the logarithm of the determinants would be stored as slow-changing parameters. The above equations have been programmed at this Center and successfully evaluated for two-dimensional sets of data.

A serious practical problem arises when the observed unknown has one or more missing or very noisy measurements. Some effort was expended on an analytical solution to this problem, and this appears in Appendix C. It has been suggested that the best that can be done is to ignore the missing

dimension entirely; that is, deal with a subspace. For the sake of practicality, we suggest that the unacceptable measure, $x_i$, be replaced by the following weighted interpolation:

$$x_i = \frac{(u_{ii}^{k})\bar{x}_i^{h^2} + (u_{ii}^{h})\bar{x}_i^{k^2}}{(u_{ii}^{k}) + (u_{ii}^{h})} \qquad (17)$$

This interpolation is an attempt to force $x_i$ into a position between the means of the classes k and h so as not to bias the quadratic form toward one class or the other. All of the parameters of Equation (17) are readily available from the sample means and sample covariance matrices.

G. Confidence. It is important, for operational purposes, to have some measure of the level of confidence in the likelihood estimate based on an unknown X. This is a very difficult task which has never been done, to our knowledge, even for the case where the multivariate normal assumption is made. There is no way of knowing, or expressing, the condition where the probability estimates are in error due to learning data which is not completely representative; i. e., does not exist in all regions of the space in which unknowns might occur. For this reason the statistical confidence measures described here are simply a function of the number of available samples. Statistical confidence is expressed as the probability that a sample from a random variable falls within a given range of the random variable. It was not possible to give a complete analytical expression of statistical confidence without making strong simplifying assumptions about the forms of the probability density functions of the active sonar targets of interest; even then, the analytical problems were severe.

We took two approaches to estimating confidence levels: (1) Monte Carlo techniques, and (2) intuitive technique with some analytical basis. The Monte Carlo approach chosen was based upon the d-variate normal fit method of probability density estimation using clusters of more or less describable shapes (ellipsoidal and rectangular). The validity of this error analysis therefore applies only to cluster analysis methods which control the shape of the resulting clusters. Where the cluster shapes are not controlled or are only partially controlled, the idea of a confidence measure for the d-variate normal fit becomes meaningless; i.e., our

confidence would always be very low. The Monte Carlo attempt was not completed due to time restrictions; however, the following procedure for analyzing the problem was defined:

1. Randomly choose a covariance matrix, $\Sigma_{d \times d}$

2. Randomly generate a number of samples, M, from $(0,\Sigma)$ using the TDATA Program.

3. Compute a sample covariance matrix, U, and sample mean $\overline{X}$ from the M samples.

4. Using Equation (11), compute $\hat{P}(X)$ for a variety of points $\{X\}$.

5. Using $\Sigma$ (the "true" covariance matrix) and a mean of zero (the "true" mean), use Equation (11) to get the "true" P(X) values.

6. Compute the relative error $(\hat{P}(X)-P(X))/P(X)$ and other error functions for each X.

7. Record: d, M, $|U|$, Q, and the error functions.

8. Repeat Steps 2-7 some number of times and then go back to Step 1.

The approach here was to treat the relative error as a random variable which was a function of the other random variables d, M, $|U|$, and Q. A FORTRAN program was written to carry out these steps, but was never completely debugged. The ultimate intent was to fit the surfaces of relative error and other error descriptions as a function of d, M, $|U|$, Q. This would have allowed an error and confidence level approximation for each X with some experimental justification. Similarly, a Monte Carlo error analysis could also be made when the cluster samples are used to generate any other probability density function under evaluation.

In the case of PDM the polynomials would be prechosen as the source probability density functions. Confidence in this case could be defined as a function of the number of learning samples, the degree of the polynomial, and the number of dimensions.

The second approach was to view the problem in a stylized way so that some helpful analysis might be applied. We made an assumption that the multivariate normal function represented the shape of the "true" probability density function well enough (most samples in the cluster near the mean, diminishing to a very few at the edges). The second assumption was that the only error was in the location of the unknown, X, with respect to X. In this case, the relative error can be approximated by

$$\text{relative error} \simeq \frac{\Delta X \cdot \text{grad } \hat{P}(X)}{\hat{P}(X)} \tag{18}$$

where $\Delta X$ denotes an error in the distance to the mean. The derivation of the gradient is given in Appendix D. To estimate $\Delta X$, one may make use of the fact that the variance of the sample mean of a normal population is 1/M of the population variance. This allows the following extension:

$$\Delta X = (|U|/M^d)^{1/2} \tag{19}$$

where $|U|$ is the determinant of the sample covariance matrix of M samples in d-dimensions. Once again, there was insufficient time to test the feasibility of this approximation on the computer, and it is included here as being suggestive of a possible direction for future work.

H. Number of Parameters. The parameters that the system has to learn using a gaussian fit for each cluster in a measurement space of d-dimensions are the d-components of the sample mean, the d-variance elements, and the $d(d - 1)/2$ covariance elements of the sample covariance matrix. This is a total of $(d^2 + 3d)/2$ parameters that would have to be retained for each cluster. Table 3 shows the number of parameters for a range of dimensions (d) and number of clusters (n).

Note that the tabulated values are equal to $n(d^2 + 3d)/2$.

TABLE 3

NUMBER OF MULTIVARIATE NORMAL LEARNING PARAMETERS

| Number of Dimensions (d) | Number of Clusters (n) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5 | 20 | 40 | 60 | 80 | 100 | 120 | 140 |
| 6 | 27 | 54 | 81 | 108 | 135 | 162 | 189 |
| 7 | 35 | 70 | 105 | 140 | 175 | 210 | 245 |
| 8 | 44 | 88 | 132 | 176 | 220 | 264 | 308 |
| 9 | 54 | 108 | 162 | 216 | 270 | 324 | 378 |
| 10 | 67 | 130 | 195 | 280 | 320 | 390 | 455 |
| 11 | 77 | 154 | 231 | 308 | 385 | 462 | 539 |
| 12 | 90 | 180 | 270 | 360 | 450 | 540 | 630 |
| 13 | 104 | 208 | 312 | 416 | 520 | 624 | 728 |
| 14 | 119 | 238 | 357 | 476 | 595 | 714 | 833 |
| 15 | 135 | 270 | 405 | 540 | 675 | 810 | 945 |
| 16 | 152 | 304 | 456 | 608 | 760 | 912 | 1064 |

1. The Polynomial Discriminant Method (PDM). The Polynomial Discriminant Method is introduced here as the exemplar of the non-parametric statistical approaches to classification. It was planned earlier in this project to compare the PDM with the cluster analysis/multivariate normal fit (CA/MNF) approach discussed in preceding sections. As with certain other portions of the investigation, there was insufficient time to implement PDM on a computer and carry out such a comparison.

The following paraphrases the important features of PDM and compares it with the cluster analysis/multivariate normal fit approach. In the PDM, the approach is to view each sample as independently representing a local parent density whose form is a spherically symmetrical normal function. The overall parent density is just the averaged sum of these functions expanded principally as a polynomial. The spread of each individual contribution, $\sigma$, can be adjusted to compensate fot the "bumpy" density which arises from small numbers of samples. The PDM worked roughly as follows:

1. The learning algorithm requires that each sample be used one at a time, so there is no need to store each sample after it has been observed. The same is true of CA/MNF, inasmuch as nothing nore than averaging is involved.

2. The algorithms for calculating the polynomial coeffieients are fairly simple, as is also the case with calculating sample means and sample covariance matrices.

3. The shape of the polynomial density function can be made as complex or simple as desired by adjusting $\sigma$, the spread parameter. With CA/MNF, the density function is a union of ellipsoidal shapes. While the

latter can get complex when there are many clusters, it is restricted as compared to the PDM.

4. When the PDM is used for classification, the surfaces of equal likelihood can be strictly linear or highly non-linear, depending on $\sigma$. Such surfaces are always second degree for CA/MNF.

5. The PDM will work with only one sample. It is necessary to have at least d samples to calculate a non-singular d x d sample covariance matrix.

6. Theoretically, the PDM does not require any preliminary cluster analysis. The CA/MNF has no meaning for unclustered data, and may still work poorly unless the cluster is more or less convex.

7. The number of polynomial coefficients required increases geometrically with the sum of the degree and number of dimensions (variables). The basis for a comparison of the number of PDM coefficients with the number of CA/MNF parameters is important enough to be developed in the following section.

J. Comparison of Number of Learning Parameters. In the case where each cluster in a sample set corresponds to a mode in the parent density, the degree of the polynomial necessary to represent the parent density must be at least one greater than the number of modes. Where r is the degree of the polynomial, n is the number of modes, and d is the number of dimensions, the number of terms in the polynomial is:

$$\text{number of terms} = \frac{(r + d)!}{r!\ d!} \tag{20}$$

$$= \frac{(n + 1 + d)!}{(n + 1)!\ d!} \tag{21}$$

Table 4 gives the number of terms for a range of n and d for comparison with Table 3. The proper interpretation of the tables is not that the number of terms in PDM is prohibitively large, as one might be tempted to think. Note that for the case where there is only one mode, the number of parameters is virtually the same as for the CA/MNF. We feel that it would be possible to get the best results by combining cluster analysis with PDM; i. e., fit polynomial density functions to clusters rather than attempt to use the entire set of learning data to generate a single polynomial.

**TABLE 4**

MAXIMUM NUMBER OF POLYNOMIAL TERMS

| Number of Dimensions, d | Number of Modes, n | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 5 | 21 | 56 | 126 | 252 | 462 | 792 | 1287 |
| 6 | 28 | 84 | 210 | 462 | 924 | 1716 | 3003 |
| 7 | 36 | 120 | 330 | 792 | 1716 | 3432 | 6435 |
| 8 | 45 | 165 | 495 | 1287 | 3003 | 6435 | 12,870 |
| 9 | 55 | 220 | 715 | 2002 | 5005 | 11,440 | 24,310 |
| 10 | 66 | 286 | 1001 | 3003 | 8008 | 19,448 | 43,758 |
| 11 | 78 | 364 | 1365 | 4368 | 12,376 | 31,824 | 75,582 |
| 12 | 91 | 455 | 1820 | 6188 | 18,564 | 50,388 | 125,970 |
| 13 | 105 | 560 | 2380 | 8568 | 27,132 | 77,520 | 203,490 |
| 14 | 120 | 680 | 3060 | 11,628 | 38,760 | 116,280 | 319,770 |
| 15 | 136 | 816 | 3876 | 15,504 | 54,264 | 170,544 | 490,314 |
| 16 | 153 | 969 | 4845 | 20,349 | 74,613 | 245,157 | 735,471 |

## IV. TARGET CLASSIFICATION SUBSYSTEM DESIGN

Figure 1, a functional block diagram, depicts the principal elements of a target classification subsystem as it might be implemented within the data processing and display complex of an advanced active sonar system. This subsystem design incorporates the ideas and techniques which appear to be most promising, and at the same time are compatible with total system constraints. For convenience, the elements of the target classification subsystem are grouped into the three subfunctions of preprocessing, learning and adaptation, and classification. These subfunctions are discussed in order.

A. Preprocessing. The preprocessing subfunction embraces the digital processes that reduce the inputs from the target detection and tracking programs to vector representations useful to the other subfunctions. The measurement extraction programs produce a finite set of target measurements for each new look at a target. Varying degrees of extraction and processing will be required to obtain this set of measurements. For example, target speed will be available directly from the tracking program, whereas target aspect angle and depth will require some computations. The measurement extraction program is designed to reduce the quantity of target data available and represent these data in a manner which emphasizes the features that distinguish targets of interest.

When the desired measurements are extracted they will, in general, be in different units. Normalization of the measurements can be accomplished by dividing by their respective standard deviations, a procedure which is equivalent to a scalar transformation. After scaling, it may be desirable

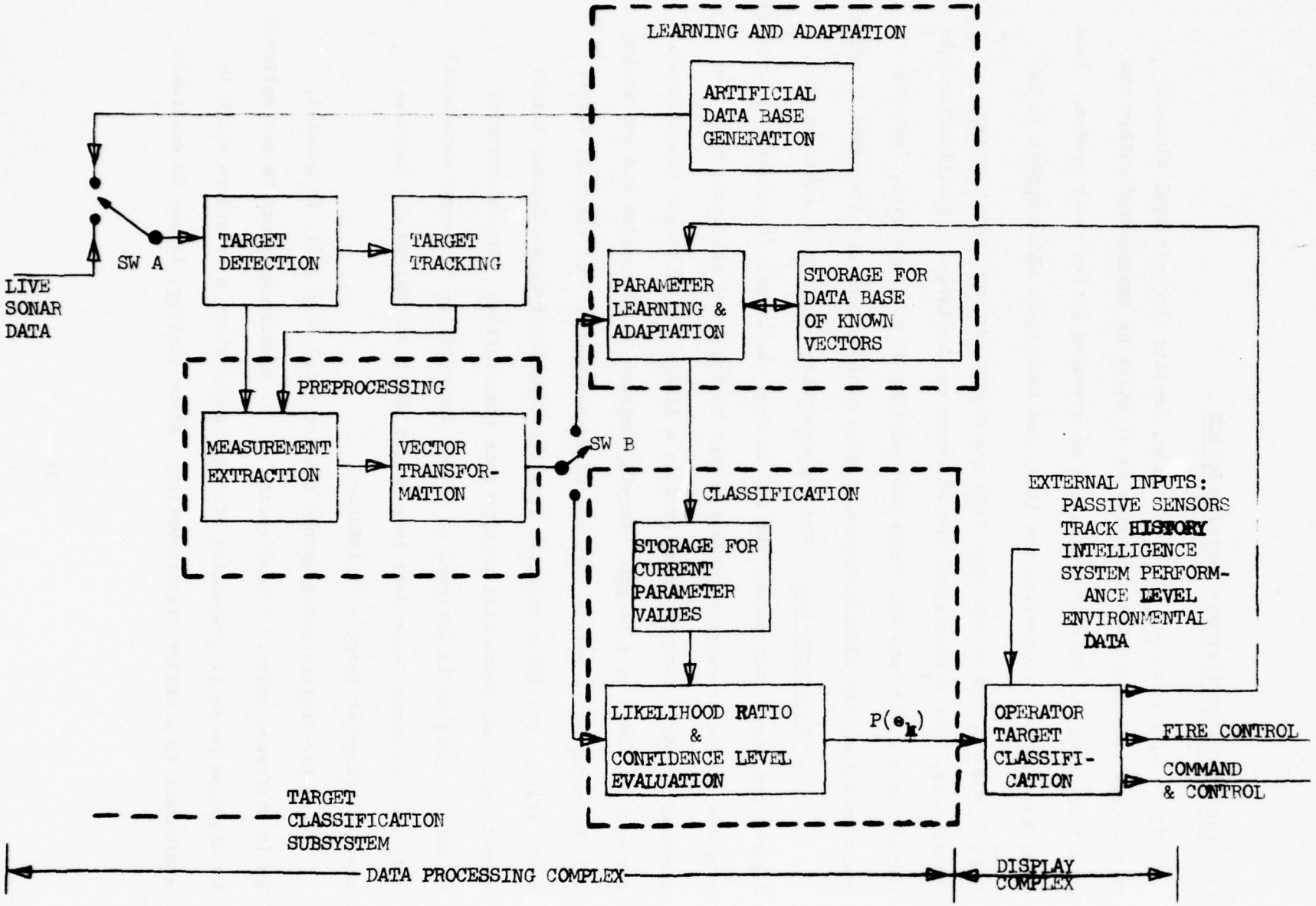



FIGURE 1. ADVANCED ACTIVE SONAR SYSTEM DATA PROCESSING AND DISPLAY COMPLEX INCLUDING TARGET CLASSIFICATION SUBSYSTEM.

either to transform these vectors into a more propitious coordinate system by performing an eigenvalue analysis, or to reduce the dimensionality by doing a suitable mapping. The procedures which take the extracted normalized measurements and produce the vectors {X} to be learned or classified are termed "vector transformations" in the figure.

B. Learning and Adaptation. Learning (Switch B "UP" of Fig. 1) occurs when the subsystem is exposed either to artificial data (Switch A "UP") or to real sonar data (Switch A "DOWN"). In either case, when a vector is presented for learning, its class membership must also be given. This new vector is included in the data base for known vectors prior to re-initialization of the cluster analysis routines which recompute (learn) new parameter values or adapt existing ones. Note that the idea of adaptation also includes the special case of one-shot learning; this is essentially adaptation from a state of complete ignorance. In this sense, "learning" and "adaptation" can be used interchangeably.

The element referred to in Fig. 1 as Parameter Learning and Adaptation is completely general because the learned parameters could be coefficients of a polynomial, the parameters of a d-variate normal distribution, or the parameters of any other function that describes class probability densities. Whatever their form, the most up-to-date values for the parameters are stored in high-speed core memory for immediate use in evaluating the likelihood ratios and confidence levels for unknown vectors.

The function of storing a data base of known vectors is important as it retains the information required to perform meaningful cluster analysis

on a background basis. Also, vectors from tracks that are as yet unclassified must be retained until the operator classifies them. These vectors could be stored temporarily in the data base as unknowns. The allocation of space for these vectors would depend on operational data rates and the amount of storage available.

C. Classification. In a completely automatic classification subsystem, the conditional probability of membership in the several classes could be estimated using the current parameter values and then perhaps combining with cost functions and a priori probabilities to produce a classification decision. The absence of reliable cost and a priori information in actual operational situations are among the reasons why the ultimate responsibility for target classification still remains with the sonar operator. In Figure 1 the automatic output of the classification subfunction is depicted as an independent input to the sonar operator. This operator must continually exercise his own judgment based on the total of the information received from the target classification subsystem, $P(\theta_k|X)$, the processed outputs from the passive sensors, passive and active track histories, intelligence information, environmental data, and system performance level for the sonar unit. The classification decisions of the sonar operator are then sent to the personnel responsible for fire control and command and control decisions.

When an operator classifies an active track, all samples from that track will be fed back automatically to the target classification subsystem and used to improve the previously learned parameter values by the adaptation process. This will be done as background processing and will be effected only when all immediate operational needs have been met.

## V. RECOMMENDATIONS

During the course of this investigation, certain additional approaches of considerable promise were isolated but were not completely evaluated due to lack of time. If further work of an exploratory nature is contemplated for automatic classification, the authors feel that the following recommended approaches should be given consideration:

1. The Polynomial Discriminant Method of Dr. Donald F. Specht (Stanford Electronics Laboratories) appears to have several strong points in its favor. We feel that Dr. Specht should be contacted to learn the current status of his work on the PDM. Some of his work was done under Navy contract on the POSEIDON project and should be available to the Center. Also, it would be worth-while to know of improvements to his PDM and of any differences between theory and the computer implementation. Dr. Specht's investigation has been in progress for several years and the programs developed in its course undoubtedly reflect considerable refinement. It would seem desirable to acquire these programs directly and convert them for use on the Center's computers.

2. Work should be continued on cluster analysis until a method emerges that is significantly superior and more reliable than the other leading contenders. What is needed is not an "ultimate " method but the best among at least three promising methods. Those which currently appear most promising are:

   a. Gradient Method

   b. ISODATA

   c. An adaptation of "hill climbing procedures" to find the maxima of a PDM polynomial.

3. An effort should be made to introduce man-machine interaction into the development and evaluation of classification methods. This could be accomplished by direct CRT display of data such as is done in the PROMENADE system.[9] Or the dynamic results of cluster isolation, learning, and adaptation procedures could be displayed on specialized formats. Ideally, the investigator would be able to modify the controlling parameters of these procedures on-line by console operator inputs. The programming of these procedures would be for the general d-dimensional data with the display representing a transformation into two dimensions or a two-dimensional subspace.

There are three important advantages that would be derived from the development of this on-line display system: (1) the development, debugging, and evaluation of learning, adaptation, and probability estimation procedures would be greatly speeded up, (2) real data could be examined in detail and at length to determine which combinations of measurements are most valuable for distinguishing the important active sonar targets, and (3) the display would provide the means to demonstrate visually the strongly intuitive concepts of automatic classification. The display would serve to reduce the handicaps of technical terminology and multi-dimensionality by providing a dynamic representative of the classification procedures.

4. The prototype classification subsystem described in Section IV of this note should be implemented in FORTRAN and perhaps JOVIAL. The choice of these particular languages would facilitate a comparison with other attempts at automatic classification subsystems designed under Government and military auspices.

5. A method is needed for expressing confidence levels in a manner which is both easy to compute and meaningful in an operational situation. This means the development of a method which provides acceptable results but does not consume computational time or space out of proportion to its usefulness. The following is a recommended approach for developing such a measure of confidence:

a. Generate data which will be representative of the operational situation. This data would be generated by a statistical model where the parent distributions, P(X)'s, would be known.

b. Reduce the data by the cluster analysis method chosen for operational implementation.

c. For each cluster estimate the probability density, $\hat{P}(X)$, of the parent distribution.

d. Compute the error $E = \hat{P}(X) - P(X)$ and chosen functions of the error such as $E^2$, $|E|$, etc., over a range of points in the data space.

e. Analyze the results of the error computations from several trials over a wide range of the independent variables such as number of samples, number of dimensions, parameters of the parent probability function, etc.

f. Relate the error functions to the independent variables. A preliminary approach would be to plot the error functions as contours for pairwise combinations of independent variables. It may be possible from this analysis to eliminate the least sensitive variables.

g. Formulate the relationship between error and the selected independent variables in terms of a measure of confidence level.

## VI. CONCLUSIONS

The conclusions which can be drawn from the work expended on this task may be stated briefly as follows:

1. Existing statistical techniques or extensions of these techniques are presently available and applicable to the problem of automatic target classification by computer in active sonar systems.

2. Implementation of these techniques into a workable target classification subsystem such as that described in Section IV does appear to be feasible in view of operational and system constraints.

3. An effort should be initiated to refine the available classification techniques further and to incorporate them into an operating subsystem which is capable of demonstrating active sonar target classification on a real-time basis.

## REFERENCES


1. Roese, J. A., Computer Target Classification and Threat Evaluation in Advanced Active Sonar Systems, Naval Undersea Research and Development Center Technical Note 543, May 1971.

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

3. Nilsson, N. J., Theoretical and Experimental Investigations in Trainable Pattern-Classifying Systems, Rome Air Development Center Technical Report 65-257, September 1965.

4. Sebestyen, G. S., and Edie, J., "An Algorithm for Non-Parametric Pattern Recognition", IEEE Transactions on Electronic Computers, Vol. EC 15 No. 6, December 1966.

5. Butler, G. A., "A Vector Field Approach to Cluster Analysis", Pattern Recognition, Vol. I, pp 291-299, 1969.

6. Allais, D. C., The Selection of Measurements for Prediction, Stanford Electronics Laboratories Technical Report 64-115, November 1964.

7. Nilsson, N. J., Learning Machines: Foundations of Trainable Pattern Classifying Systems, McGraw-Hill, 1965.

8. Anderson, T. W., Introduction to Multivariate Statistical Analysis, John Wiley & Sons, 1958.

9. Hall, D. J., PROMENADE - An On-Line Pattern-Recognition System, Final Report, Contract AF 30(602)-4196, Stanford Research Institute Project 6004m August 1967.

## APPENDIX A

## A MEASURE OF TARGET RELATIVE IMPORTANCE

Suppose that we obtain a single ping estimate of the probability that a target belongs to a hostile class , $P(\Theta_h)$, and a normalized measure of confidence in that probability estimate, $0 \leq C \leq 1$. We would like to define a function, $G[P(\Theta_h),C]$, to evaluate the "importance" of the target that satisfies the following criteria:

1. When $C = 0$, $G[P(\Theta_h),C] = k$. When we have no confidence in the probability estimate, all targets are equally important.

2. When $C = 1$ and $P(\Theta_h) = 1$, G is a maximum. The most important target is the one that we are positively sure is hostile.

3. When $C = 1$ and $P(\Theta_h) = 0$, G is a minimum. The least important target is the one that we are positively sure in not hostile.

A function which satisfies the above requirements is

$$G = k + (P - k)C, \quad 0 \leq k \leq 1$$

| Criterion | G | P | C |
|---|---|---|---|
| 1 | k | any | 0 |
| 2 | max | 1 | 1 |
| 3 | min | 0 | 1 |

The choice of k determines how important the lack of confidence really is. This would suggest that k could be adjusted dynamically; that is, be kept small if C is frequently small. The objective here would be to make G sensitive to either P or C depending upon which variable is better suited to enable targets to be ranked according to their importance.

40-Blank

## APPENDIX B

### MEASURES OF CLUSTER-SEEKING PERFORMANCE

Well Separated Clusters. In the case where there are N well-separated or "true" clusters $C_1, C_2, \ldots, C_i, \ldots, C_N$ which are discerned as M "apparent" clusters $K_1, \ldots, K_j, \ldots, K_M$, the number of samples associated with the $i^{th}$ true cluster and the $j^{th}$ apparent cluster, $f_{ij}$, can be recorded in a frequency matrix $F = (f_{ij})_{N \times M}$. The mutual information, I, has a maximum value in this application when all samples are correctly associated with the original N clusters, and it is zero when the apparent clusters are totally confused; i. e., when $\sum_{j=1}^{M} f_{ij}$ is proportional to the number of samples in each $C_i$.

$$I = \frac{H(C) + H(K) - H(C * K)}{H(C)} \tag{B-1}$$

$$H(C) = \sum_{i=1}^{N} F_i \ln F_i \; ; \; F_i = \sum_{j=1}^{M} f_{ij} \tag{B-2}$$

$$H(K) = \sum_{j=1}^{M} F_j \ln F_j \; ; \; F_j = \sum_{i=1}^{N} f_{ij} \tag{B-3}$$

$$H(C * K) = \sum_{i=1}^{N} \sum_{j=1}^{M} f_{ij} \ln f_{ij} \tag{B-4}$$

The purpose of using H(C) in the denominator is to normalize $0 \leq I \leq 1$. A cluster-seeking method should score close to L if its logic has been programmed correctly and its logic is correct to begin with.



42-Blank

In the general case, randomly generated clusters will overlap in varying degrees so that the system of apparent clusters may not always be compared to the true clusters as in the well-separated case. Therefore, an additional effort would have to be made to reward the resolution of overlapping clusters and take for granted the discernment of isolated clusters. A convenient measure of the separation between two clusters is Sebestyen's interset distance, $S[(X)_m, (Y)_n]$, a mean squared distance computed as below:

$$S = \frac{\sum_{j=1}^{m} \sum_{k=1}^{n} \sum_{i=1}^{d} (x_{ij} - y_{ik})^2}{m.n} \qquad \text{(B-5)}$$

$$= \frac{\sum_{i=1}^{d} x_i^2}{m} + \frac{\sum_{i=1}^{d} y_i^2}{n} - 2 \frac{\sum_{i=1}^{d} x_i y_i}{m.n}$$

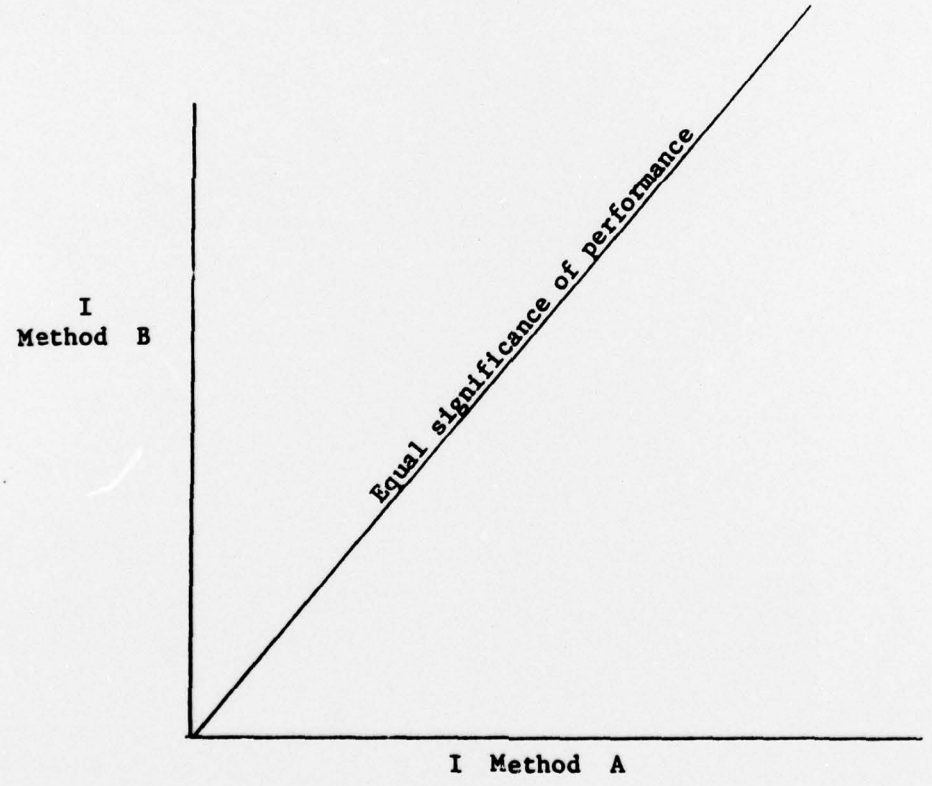
We may now define a normalized measure of the "resemblance" R between two clusters:

$$R = a^{-s} \; ; \; 0 \leq R \leq 1 \; ; \; a > 0 \qquad \text{(B-6)}$$

Designating the average resemblance of the $i^{th}$ true cluster to the other true clusters by $R_{ij}$, we can define a measure of significant performance

$$I = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} f_{ij} R_{ij}}{T} \; ; \quad 0 \leq I \leq 1 \qquad \text{(B-7)}$$

where T is the total number of samples. This function is not at all similar to the I suggested for the well-separated case. Here, the apparent clusters may agree with the true clusters perfectly, but achieve I = 0 if the true clusters are trivially well separated. High scores can only be achieved when overlapping true clusters are correctly identified. The indicated use of this general measure is to compare two methods, perhaps by a scatter diagram as illustrated below.



## APPENDIX C

### MISSING MEASUREMENTS

In the case where all d-dimensions are observed, use is made of a stored determinant of the sample covariance matrix and its inverse to compute

$$P(X) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|U|^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j u_{ij}\right) \qquad \text{(C-1)}$$

where $u_{ij}$ is an element of $U^{-1}$ and $x_i$ is distance of the $i^{th}$ measurement with respect to the $i^{th}$ sample mean. However, if measurements are missing, they can be ignored in the computation of the exponent; namely, the program will omit terms for which i or j corresponds to a missing measurement. The term $(2\pi)^{d/2}$ can simply be obtained from a table indexed by d. The new determinant, however, does present a computational problem. There appear to be three possible approaches:

1. Compute the determinant of the remaining matrix when the rows and columns corresponding to the missing measures are removed. This is straightforward, but time-consuming and potentially redundant.

2. Reduce the "full" determinant by making use of stored minors. This works well for one missing measure, but becomes awkward thereafter. It may still be the most expedient approach in the last analysis. If k is the only missing measure, the desired determinant is

$$\text{minor of } u_{kk} = \frac{\text{cofactor } u_{kk}}{(-1)^{k+k}} \tag{C-2}$$

$$= \text{cofactor } u_{kk}$$

$$|U| = \sum_{j=1}^{d} u_{kj} \text{ cofactor}_{kj} \tag{C-3}$$

$$= \sum_{j \neq k}^{d} u_{kj} \text{ cofactor}_{kj} + u_{kk} \text{ cofactor}_{kk}$$

$$\text{minor of } u_{kk} = \text{cofactor } u_{kk} = \frac{|U| - \left( \sum_{j \neq k}^{d} u_{kj} \text{ cofactor}_{kj} \right)}{(u_{kk})} \tag{C-4}$$

The terms in parentheses could easily be stored for each measure.

3. Estimate the new determinant as a function of $|U|$ and correlations among the variables. This approach has not been explored, but appears to be promising.

## APPENDIX D

## GRADIENT OF MULTIVARIATE NORMAL DISTRIBUTION

With sample size and number of dimensions held constant, it is reasonable to assume that the error in the estimated probability will be proportional to the gradient at the point $X = (x_1,\ldots,x_d)$, that is, the maximum rate at which $P(X)$ is changing at $X$.

$$P(X) = \frac{1}{(2\Pi)^{d/2}} \frac{1}{|U|^{1/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j u_{ij}\right) \qquad \text{(D-1)}$$

$$= C \exp\left(-\frac{1}{2} Q\right)$$

To differentiate with respect to the $k^{th}$ variable (regarding the remaining d-1 variables as constant), Q can be rewritten as

$$Q = 2x_k \sum_{j\neq k}^{d} x_j u_{kj} + x_k^2 u_{kk} + \sum_{i\neq k}^{d} \sum_{j\neq k}^{d} x_i x_j u_{ij} \qquad \text{(D-2)}$$

The partial derivative with respect to $x_k$ is obtained in the following steps:

$$\frac{dP(X)}{dQ} = -\frac{1}{2} C \exp\left(-\frac{1}{2}Q\right) \qquad \text{(D-3)}$$

$$\frac{dQ}{dx_k} = 2u_{kk}x_k + 2 \sum_{j \neq k}^{d} x_j u_{kj} \tag{D-4}$$

$$\frac{dP(X)}{dx_k} = \frac{dP(X)}{dQ} \cdot \frac{dQ}{dx_k} \tag{D-5}$$

$$= -C(u_{kk}x_k + \sum_{j \neq k}^{d} x_j u_{kj}) \exp\left(-\frac{1}{2} Q\right)$$

$$= -C(\sum_{j=1}^{d} x_j u_{kj}) \exp\left(-\frac{1}{2} Q\right)$$

Finally, the absolute value of the gradient is the root of the sum of the squares of the d partial derivatives.

$$\left[\frac{dP(X)}{dx_k}\right]^2 = C^2 \exp(-Q)\left(\sum_{j=1}^{d} x_j u_{kj}\right)^2 \tag{D-6}$$

$$\sum_{k=1}^{d}\left[\frac{dp(X)}{dx_k}\right]^2 = C^2 \exp(-Q) \sum_{k=1}^{d}\left(\sum_{j=1}^{d} x_j u_{kj}\right)^2 \tag{D-7}$$

$$= C^2 \exp(-Q) \sum_{j=1}^{d} x_j \left(\sum_{k=1}^{d} u_{kj}\right)^2$$

$$\text{Grad} = C \quad \exp\left(-\frac{Q}{2}\right) \left[ \sum_{j=1}^{d} x_j^2 \left( \sum_{k=1}^{d} u_{kj} \right) \right]^{1/2} \qquad \text{(D-8)}$$

$$= P(X) \left[ \sum_{j=1}^{d} x_j^2 \left( \sum_{k=1}^{d} u_{kj}^2 \right) \right]^{1/2}$$

The gradient is therefore the product of two terms, one of which is P(X). The gradient thus goes to zero as P(X) goes to zero. The gradient is also zero when X = 0. It can be verified that the gradient is at a maximum at one standard deviation from the mean in any direction.