DOCUMENT ROOM, [illegible] DOCUMENT ROOM 36-412
RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE 39, MASSACHUSETTS, U.S.A.

#3

CLASSIFICATION DECISIONS IN PATTERN RECOGNITION

GEORGE S. SEBESTYEN

LOAN COPY only

TECHNICAL REPORT 381

APRIL 25, 1960

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the U. S. Army (Signal Corps), the U.S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039-sc-78108, Department of the Army Task 3-99-20-001 and Project 3-99-00-000.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 381 April 25, 1960

CLASSIFICATION DECISIONS IN PATTERN RECOGNITION

George S. Sebestyen

This report is based on a thesis, entitled "On Pattern Recognition with Application to Silhouettes," submitted to the Department of Electrical Engineering, M. I. T., August 24, 1959, in partial fulfillment of the requirements for the degree of Doctor of Science. A portion of the work represents an extension of the thesis and was carried out at Melpar, Inc., Watertown, Massachusetts, under Contract AF30(602)-2112.

Abstract

The basic element in the solution of pattern-recognition problems is the requirement for the ability to recognize membership in classes. This report considers the automatic establishment of decision criteria for measuring membership in classes that are known only from a finite set of samples. Each sample is represented by a point in a suitably chosen, finite-dimensional vector space in which a class corresponds to a domain that contains its samples. Boundaries of the domain in the vector space can be expressed analytically with the aid of transformations that cluster samples of a class and separate classes from one another. From these geometrical notions a generalized discriminant analysis is developed which, as the sample size goes to infinity, leads to decision-making that is consistent with the results of statistical decision theory.

A number of special cases of varying complexity are worked out. These differ from one another partly in the manner in which the operation of clustering samples of a class and the separation of classes is formulated as a mathematical problem, and partly in the complexity of transformations of the vector space which is permitted during the solution of the problem. The assumptions and constraints of the theory are stated, practical considerations and some thoughts on machine learning are discussed, and an illustrative example is given for the automatically learned recognition of spoken words.

TABLE OF CONTENTS

I. INTRODUCTION

As the advances of modern science and technology furnish the solutions to problems of increasing complexity, a feeling of confidence is created in the realizability of mathematical models or machines that can perform any task for which a specified set of instructions for performing the task can be given. There are, however, problems of long-standing interest that have eluded solution, partly because the problems have not been clearly defined, and partly because no specific instructions could be given on how to reach a solution. Recognition of a spoken word independently of the speaker who utters it, recognition of a speaker regardless of the spoken text, threat evaluation, the problem of making a medical diagnosis, and that of recognizing a person from his handwriting are only a few of the problems that have remained largely unsolved for the above-mentioned reasons.

All of these problems in pattern recognition, however different they may seem, are united by a common bond that permits their solution by identical methods. The common bond is that the solution of these problems requires the ability to recognize membership in classes, and, more important, it requires the automatic establishment of decision criteria for measuring membership in each class.

The purpose of this report is to consider methods of automatically establishing decision criteria for classifying events as members of one or another of the classes when the only knowledge about class membership is from a finite set of their labeled samples.

We shall consider the events represented by points or vectors in an N-dimensional space. Each dimension expresses a property of the event, a type of statement that can be made about it. The entire signal that represents all of the information available about the event is a vector $V = (v_1, v_2, \ldots, v_n, \ldots, v_N)$, the coordinates of which have numerical values that correspond to the amount of each property present in the event. In this representation, the sequence of events belonging to the same category corresponds to an ensemble of points scattered within some region of the signal space.

The concept playing a central role in the theory that will be described is the notion that the ensemble of points in signal space that represents a set of nonidentical events belonging to a common category must be close to each other, as measured by some — as yet — unknown method of measuring distance, since the points represent events that are close to each other in the sense that they are members of the same category. Mathematically speaking, the fundamental notion underlying the theory is that similarity (closeness in the sense of belonging to the same class or category) is expressible by a metric (a method of measuring distance) by which points representing examples of the category we wish to recognize are found to lie close to each other.

To give credence to this idea, consider what we mean by the abstract concept of a class. According to one of the possible definitions, a class is a collection of things that have some common properties. By a modification of this thought, a class could be characterized by the common properties of its members. A metric by which points

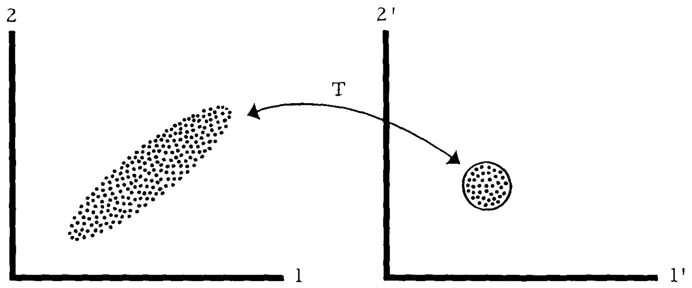


Fig. 1. Clustering by transformation.

representing samples of a class are close to each other must therefore operate chiefly on the common properties of the samples and must ignore, to a large extent, those properties not present in each sample. As a consequence of this argument, if a metric were found that called samples of the class close, somehow it would have to exhibit their common properties.

In order to present this fundamental idea in a slightly different way, we can state that a transformation on the signal space that is capable of clustering the points representing the examples of the class must operate primarily on the common properties of the examples. A simple illustration of this idea is shown in Fig. 1, where the ensemble of points is spread out in signal space (only a two-dimensional space is shown for ease in illustration) but a transformation T of the space is capable of clustering the points of the ensemble. In the example above, neither the signal's property represented by coordinate 1 nor that represented by coordinate 2 are sufficient to describe the class, for the spread in each is large over the ensemble of points. Some function of the two coordinates, on the other hand, would exhibit the common property that the ratio of the value of coordinate 2 to that of coordinate 1 of each point in the ensemble is nearly one. In this specific instance, of course, simple correlation between the two coordinates would exhibit this property; but in more general situations simple correlation will not suffice.

If the signal space shown in Fig. 1 were flexible (like a rubber sheet), the transformation T would express the manner in which various portions of the space must be stretched or compressed in order to bring the points together most closely.

Although thinking of transformations of the space is not as general as thinking about exotic ways of measuring "distance" in the original space, the former is a rigorously correct and easily visualized analogy for many important classes of metrics.

Mathematical techniques will be developed to find automatically the "best" metric or "best" transformation of given classes of metrics according to suitable criteria that establish "best."

As in any mathematical theory, the theory that evolved from the preceding ideas is based on certain assumptions. The first basic assumption is that the N-dimensional signal space representation of events exemplifying their respective classes is sufficiently complete to contain information about the common properties that serve to characterize the classes. The significance of this assumption is appreciated if we consider, for example, that the signal space contains all of the information that a black-and-white television picture could present of the physical objects making up the sequence of events which constitute the examples of a class. No matter how ingenious are the data-processing schemes that we might evolve, objects belonging to the category "red things" could not be identified because representation of the examples by black-and-white television simply does not contain color information. For any practical situation one must rely on engineering judgment and intuition to determine whether or not the model of the real world (the signal space) is sufficiently complete. Fortunately, in most cases, this determination can be made with considerable confidence.

A second assumption states the class of transformations or the class of metrics within which we look for the "best." This assumption is equivalent to specifying the allowable methods of stretching or compressing the signal space within which we look for the best specific method of deforming the space. In effect, an assumption of this type specifies the type of network (such as active linear networks) to which the solution is restricted.

The third major assumption is hidden in the statement that we are able to recognize a "best" solution when we have one. In practice, of course, we frequently can say what is considered a good solution even if we do not know which is the "best." The criterion by which the quality of a metric or transformation is judged good is thus one of the basic assumptions.

Within the constraints of these assumptions, functions of the signal space and the known, labeled sequence of events that permit the separation of events into their respective categories may be found.

Throughout this report essentially all of the mathematics is developed in simple algebraic form, with only occasional use of matrix notation in places where its use greatly simplifies the symbolism. Insistence on the algebraic form sometimes results in the loss of elegance and simplicity of solutions. But it is felt that the ease of transition from algebra to a computer program is so significant that the loss in the aesthetic appeal of this material to a mathematician must be risked. While the mathematics is thus arithmetized and computer-oriented for practical reasons, we must not lose sight of the broader implications suggested to those who are well versed in communication theory. It is a valuable mental exercise to interpret what is presented here from the point of view of communication theory. To help cross the bridge, at least partially, we can say that passing from the discrete to the continuous, from sum to integral, from dot product to correlation, and from transformation to passage through a linear network (convolution) is valid in all cases in the material contained in this report. The sets of vectors or events are sample functions of a random process, and the metrics obtained are equivalent to different error criteria. The Euclidean metric, for instance, is the mean-square error. The application of other metrics developed in most of this report is equivalent to using the mean-square-error criterion after the signal's passage through a linear network. However, some other metrics developed are error criteria that cannot be equated to a combination of filtering and mean-square-error measurement.

The material presented here is organized in essentially two parts. In Sections II and III a special form of the theory is presented in some detail so that fundamental concepts and the mechanical details of the mathematical developments could be introduced. The highlights of these two sections are the development of the notions of similarity, feature weighting, and the elementary ideas of machine learning. Toward the end of Section III, some practical considerations are pursued, primarily to estimate the storage requirements of a machine that would implement the numerical calculations.

Discussion of the second part of the material opens in Section IV with the continued application of the linear methods cited in earlier sections to the problem of clustering events that belong to the same category while separating them from those that belong to other categories. Several optimization criteria are considered, and the solutions are derived. The methods of applying the nonlinear methods within the framework of the ideas of this classificatory analysis are discussed in Section V, which also incorporates miscellaneous ideas, some remarks about the limitations of the results, and the direction that might be taken by continuing research. The brief final discussion of some of the important aspects of pattern recognition in Section VI places this work in its proper perspective and relates it to other work in the field. Appendix A discusses a network analog for solving eigenvalue problems rapidly — the most time-consuming mathematical operation encountered in implementing the techniques described. In Appendix B the relationship between decision theory and the theory developed here is explored. Appendix C contains a discussion of the numerical application of the special form of the theory to an example in the recognition of spoken words. Appendix D establishes a further connection between the theory developed here and the classical problem of statistical hypothesis testing.

II. A SPECIAL THEORY OF SIMILARITY

2.1 SIMILARITY

The central problem of pattern recognition is viewed in this work as the problem of developing a function of a point and a set of points in an N-dimensional space in order to partition the space into a number of regions that correspond to the categories to which the known set of points belongs. A convenient and special – but not essential – way of thinking about this partitioning function is to consider it as being formed from a set of functions, one for each category, where each function measures the "likelihood" with which an arbitrary point of the space could best fit into the particular function's own category. In a sense, each function measures the similarity of an arbitrary point of the space to a category, and the partitioning function assigns the arbitrary point to that category to which the point is most similar. (Although the term "likelihood" has an already well-defined meaning in decision theory, it is used here in a qualitative way

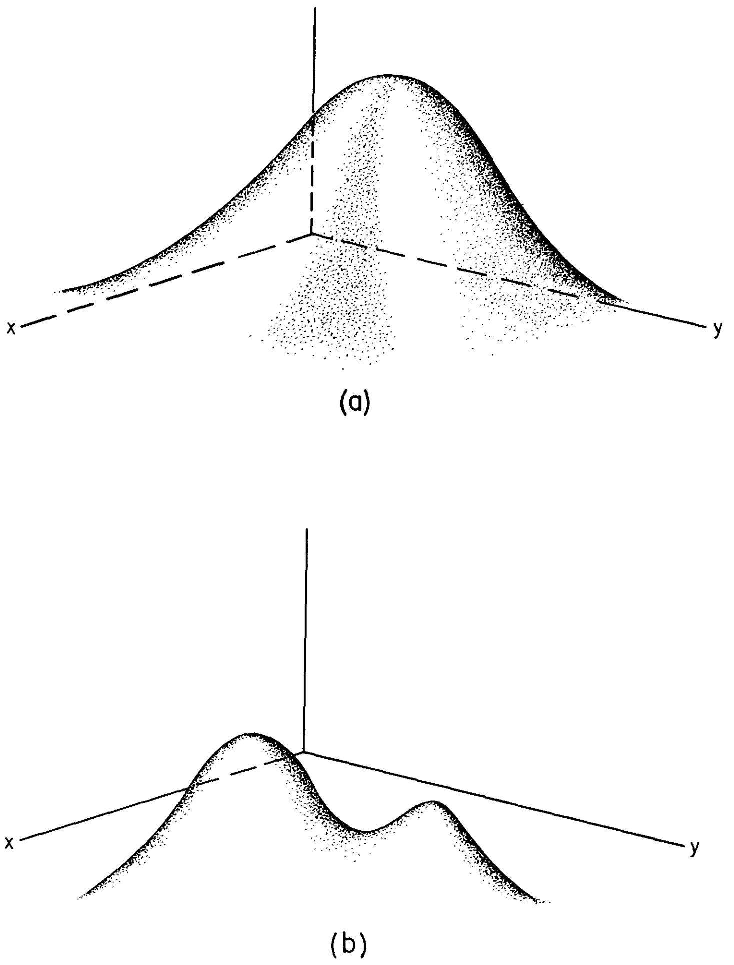


Fig. 2. Likelihood of membership in two categories: (a) Category 1; (b) Category 2.

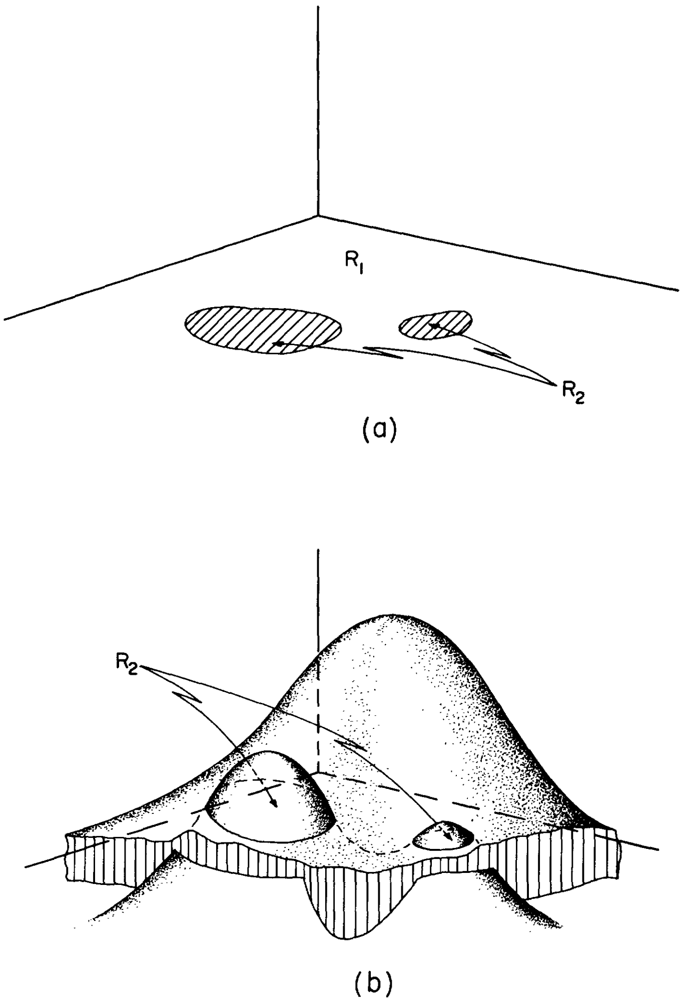


Fig. 3. Classification by maximum likelihood ratio.

to emphasize the similarity between fundamental ideas in decision theory and in the theory that is described here.)

The foregoing concept of partitioning the signal space is illustrated in Fig. 2, where the signal space has two dimensions and the space is to be partitioned into two categories. In Fig. 2a, the height of the surface above the x-y plane expresses the likelihood that a point belongs to Category 1, while that of the surface in Fig. 2b expresses the likelihood that the point belongs to Category 2. The intersection between the two surfaces, shown in Figs. 3a and 3b, marks the boundary between Region 1, where points are more likely to belong to Category 1 than to Category 2, and Region 2, where the reverse is true.

For each category of interest a set of likelihood ratios can be computed that expresses the relative likelihood that a point in question belongs to the category of interest rather than to any of the others. From the maximum of all likelihood ratios that correspond to a given point, we can infer the category to which the point most likely belongs.

The reader will recognize the idea of making decisions based on the maximum likelihood ratio as one of the important concepts of decision theory. The objective of the preceding discourse is, therefore, simply to make the statement that once a function measuring the likelihood that a point belongs to a given category is developed, there is at least one well-established precedent for partitioning signal space into regions that are associated with the different categories. The resulting regions are like a template that serves to classify points on the basis of their being covered or not covered by the template. Although in the rest of this section, partitioning the signal space is based on a measure of similarity that resembles the likelihood ratio only in the manner in which it is used, it is shown in Appendix B that, in certain cases, decisions based on the measure of similarity are identical with those based on the maximum likelihood ratio.

In the first three sections of this report, a quantitative measure of similarity is developed in a special theory in which similarity is considered as a property of only the point to be compared and the set of points that belongs to the category to be learned. In later sections we shall discuss methods for letting known nonmembers of the class influence the development of measures of similarity.

In the special theory, similarity of an event P to a category is measured by the closeness of P to every one of those events $\{F_m\}$ known to be contained in the category. Similarity S is regarded as the average "distance" between P and the class of events represented by the set $\{F_m\}$ of its examples. Two things should be noted about this foregoing definition of similarity. One is that the method of measuring distance does not influence the definition. Indeed, "distance" is not understood here in the ordinary Euclidean sense; it may mean "closeness" in some arbitrary, abstract property of the set $\{F_m\}$ that has yet to be determined. The second thing to note is that the concept of distance between points, or distance in general, is not fundamental to a concept of similarity. The only aspect of similarity really considered essential is that it is a real valued function of a point and a set that allows the ordering of points according to their similarity to the set. The concept of distance is introduced as a mathematical convenience based on intuitive notions of similarity. It will be apparent later how this concept forms part of the assumptions stated in Section I as underlying the theory to be presented. Even with the introduction of the concept of distance there are other ways of defining similarity. Nearness to the closest member of the set is one possibility. This implies that an event is similar to a class of events if it is close in some sense to any member of the class. We shall not philosophize on the relative merits of these different ways of defining similarity. Their advantages and disadvantages will become apparent as this theory is developed, and the reader will be able to judge for himself which set of assumptions is most applicable under a given set of circumstances. The essential role of the definition of similarity and the choice of the class of metrics within which the optimum is sought is to define the decision rule with which membership in classes will be determined. The decision rule, of course, is not an a priori fixed rule; it contains an unknown function, the unspecified metric, which will be tailored to the particular problem to be solved. For the time being, the decision rule will remain an ad hoc rule; it will be shown later that it is indeed a sound rule.

To summarize the foregoing remarks, for the purpose of the special theory, similarity $S(P, \{F_m\})$ of a point P and a set of points $\{F_m\}$ exemplifying a class will be defined as the mean-square distance between the point P and the M members of the set $\{F_m\}$. This definition is expressed by Eq. 1, where the metric d( ) - the method of measuring distance between two points - is left unspecified.

$$S(P,\{F_m\}) = \frac{1}{M} \sum_{m=1}^{M} d^2(P,F_m) \tag{1}$$

To deserve the name metric, the function d( ) must satisfy the usual conditions stated in Eqs. 2.

$$d(A,B) = d(B,A) \qquad \text{(symmetric function)} \tag{2a}$$

$$d(A,C) \leq d(A,B) + d(B,C) \qquad \text{(triangle inequality)} \tag{2b}$$

$$d(A,B) \geq 0 \qquad \text{(non-negative)} \tag{2c}$$

$$d(A,B) = 0 \qquad \text{if, and only if, } A = B \tag{2d}$$

2.2 OPTIMIZATION AND FEATURE WEIGHTING

In the definition of similarity of Section 2.1, the mean-square distance between a point and a set of points served to measure similarity of a point to a set. The method of measuring distance, however, was left unspecified and was understood to refer to distance in perhaps some abstract property of the set. Let us now discuss the criteria for finding the "best" choice of the metric, and then apply this optimization to a specific and simple class of metrics that has interesting and useful properties.

Useful notions of "best" in mathematics are often associated with finding the extrema of the functional to be optimized. We may seek to minimize the average cost of our decisions or we may maximize the probability of estimating correctly the value of a random variable. In the problem above, a useful metric, optimal in one sense, is one that minimizes the mean-square distance between members of the same set, subject to certain suitable constraints devised to ensure a nontrivial solution. If the metric is thought of as extracting that property of the set in which like events are clustered, the mean-square distance between members of the set is a measure of the size of the cluster so formed. Minimization of the mean-square distance is then a choice of a metric that minimizes the size of the cluster and therefore extracts that property of the set in which they are most alike. It is only proper that a distance measure shall minimize distances between those events that are selected to exemplify things that are "close."

Although the preceding criterion for finding the best solution is a very reasonable and meaningful assumption on which to base the special theory, it is by no means the only possibility. Minimization of the maximum distance between members of a set is just one of the possible alternatives that immediately suggests itself. It should be pointed out that ultimately the best solution is the one which results in the largest number of correct classifications of events. Making the largest number of correct decisions on the known events is thus to be maximized and is itself a suitable criterion of optimization that will be dealt with elsewhere in this report. Since the primary purpose of this section is to outline a point of view regarding pattern recognition through a special example, the choice of "best" previously described and stated in Eq. 3 will be used, for it leads to very useful solutions with relative simplicity of the mathematics involved. In Eq. 3, F_p and F_m are the p^{th} and m^{th} members of the set $\{F_m\}$.

$$\min\left[\overline{d^2(F_p,F_m)}^{\,p,m}\right] = \min\left[\frac{1}{M(M-1)}\sum_{m=1}^{M}\sum_{p=1}^{M} d^2(F_p,F_m)\right] \qquad \text{over all choices of } d(\quad) \tag{3}$$

Of the many different mathematical forms that a metric may take, in our special theory only metrics of the form given by Eq. 4 will be considered. The intuitive notions underlying the choice of the metric in this form are based on ideas of "feature weighting," which will be developed below.

$$d(A,B) = \left[\sum_{n=1}^{N} W_n^2(a_n-b_n)^2\right]^{1/2} \tag{4}$$

In the familiar Euclidean N-dimensional space the distance between the two points A and B is defined by Eq. 5. If A and B are expressed in terms of an orthonormal coordinate system $\{\theta_n\}$, then d(A,B) of Eq. 5 can be written as in Eq. 6, where a_n and b_n, respectively, are the coordinates of A and B in the direction of θ_n.

$$d(A,B) = |A-B| \tag{5}$$

$$d(A,B) = \left[\sum_{n=1}^{N} (a_n-b_n)^2\right]^{1/2} \tag{6}$$

We must realize, of course, that the features of the events represented by the different coordinate directions θ_n are not all equally important in influencing the definition of the category to which like events belong. Therefore it is reasonable that in comparing two points feature by feature (as expressed in Eq. 6), features with decreasing significance should be weighted with decreasing weights, W_n. The idea of feature weighting is expressed by a metric somewhat more general than the conventional Euclidean metric. The modification is given in Eq. 7, where W_n is the feature weighting coefficient.

$$d(A,B) = \left[\sum_{n=1}^{N} [W_n(a_n-b_n)]^2\right]^{1/2} \tag{7}$$

It is readily verified that the above metric satisfies the conditions stated in Eq. 2 if none of the W_n coefficients is zero; if any of the W_n coefficients is zero, Eq. 2d is not satisfied.

It is important to note that the metric above gives a numerical measure of "closeness" between two points, A and B, that is strongly influenced by the particular set of similar events $\{F_m\}$. This is a logical result, for a measure of similarity between A and B should depend on how our notions of similarity were shaped by the set of events known to be similar. When we deal with a different set of events that have different similar features, our judgment of similarity between A and B will also be based on finding agreement between them among a changed set of their features.

An alternative and instructive way of explaining the significance of the class of metrics given in Eq. 4 is to recall the analogy made in Section I regarding transformations of the signal space. There, the problem of expressing what was similar among a set of events of the same category was accomplished by finding the transformation of the signal space (again, subject to suitable constraints) that will most cluster the transformed events in the new space. If we restrict ourselves to those linear transformations of the signal space that involve only scale factor changes of the coordinates and if we measure distance in the new space by the Euclidean metric, then the Euclidean distance between two points after their linear transformation is equivalent to the feature weighting metric of Eq. 4. This equivalence is shown below, where A′ and B′ are vectors obtained from A and B by a linear transformation. The most general linear transformation is expressed by Eq. 9, where a_n' is the n^{th} coordinate of the transformed vector A, and b_n' is that of the vector B.

$$A = \sum_{n=1}^{N} a_n\theta_n \quad \text{and} \quad B = \sum_{n=1}^{N} b_n\theta_n \tag{8a}$$

$$[A'] = [A][W] \qquad [B'] = [B][W] \tag{8b}$$

$$[A'-B'] = [A-B][W] \tag{8c}$$

$$[(a_1'-b_1'), (a_2'-b_2'), \dots, (a_N'-b_N')]$$

$$= [(a_1-b_1), (a_2-b_2), \dots, (a_N-b_N)] \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{bmatrix} \tag{9}$$

The Euclidean distance between A′ and B′, $d_E(A', B')$, is given in Eq. 10.

$$d_E(A', B') = \left\{ \sum_{n=1}^{N} (a_n'-b_n')^2 \right\}^{1/2} = \left\{ \sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns}(a_s-b_s) \right]^2 \right\}^{1/2} \tag{10}$$

If the linear transformation involves only scale factor changes of the coordinates, only the elements on the main diagonal of the W matrix are nonzero, and thus $d_E(A', B')$ is reduced, in this special case, to the form given in Eq. 11.

$$\text{Special } d_E(A', B') = \left[\sum_{n=1}^{N} w_{nn}^2(a_n-b_n)^2\right]^{1/2} \tag{11}$$

The preceding class of metrics will be used in Eq. 3 to minimize the mean-square distance between the set of points.

The mathematical formulation of the minimization is given in Eqs. 12a and 12b. The significance of the constraint of Eq. 12b is, for the case considered, that every weight w_{nn} is a number between 0 and 1 (the w_{nn} turn out to be positive) and that it can be interpreted as the fractional value of the features θ_n that they weight. The fractional value that is assigned in the total measure of distance to the degree of agreement that exists between the components of the compared vectors is denoted by w_{nn}.

$$\text{minimize } \overline{D^2} = \frac{1}{M(M-1)} \sum_{p=1}^{M} \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nn}^2(f_{mn}-f_{pn})^2 \tag{12a}$$

if

$$\sum_{n=1}^{N} w_{nn} = 1 \tag{12b}$$

Although the constraint of Eq. 12b is appealing from a feature-weighting point of view, from a strictly mathematical standpoint it leaves much to be desired. It does not guarantee, for instance, that a simple shrinkage in the size of the signal space is disallowed. Such a shrinkage would not change the relative orientation of the points to each other, the property really requiring alteration. The constraint given in Eq. 13, on the other hand, states that the volume of the space is constant, as if the space were filled with an incompressible fluid. Here one merely wishes to determine the kind of rectangular box that could contain the space so as to minimize the mean-square distance among a set of points imbedded in the space.

$$\prod_{n=1}^{N} w_{nn} = 1 \tag{13}$$

The minimization problem with both of these constraints will be worked out in the following equations; the results are quite similar.

Interchanging the order of summations and expanding the squared expression in Eq. 12a yield Eq. 14, where it is recognized that the factor multiplying w_{nn}^2 is the variance of the coefficients of the θ_n coordinate. Minimization of Eq. 14 under the constraint of Eq. 12b yields Eq. 15, where ρ is an arbitrary constant. Imposing the constraint of Eq. 12b again, we can solve for w_{nn}, obtaining Eq. 16.

$$\overline{D^2} = \frac{M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left[\frac{1}{M}\sum_{m=1}^{M} f_{mn}^2 + \frac{1}{M}\sum_{p=1}^{M} f_{pn}^2 - 2\left(\frac{1}{M}\sum_{m=1}^{M} f_{mn}\right)\left(\frac{1}{M}\sum_{p=1}^{M} f_{pn}\right)\right] \tag{14a}$$

$$\overline{D^2} = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left(\overline{f_n^2} - \overline{f_n}^2 \right) = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 \tag{14b}$$

$$\left[w_{nn} \sigma_n^2 - \rho \right] = 0 \qquad n = 1, 2, \ldots, N \tag{15}$$

$$w_{nn} = \frac{\rho}{\sigma_n^2} = \frac{1}{\sigma_n^2 \sum_{p=1}^{N} \frac{1}{\sigma_p^2}} \tag{16}$$

(Note that this result is identical with that obtained in determining the coefficients of combination in multipath combining techniques encountered in long-distance communications systems.)

That the values of w_{nn} so found are indeed those that minimize $\overline{D^2}$ of Eq. 12a can be seen by noting that $\overline{D^2}$ is an elliptic paraboloid in an N-dimensional space and the constraint of Eq. 12b is a plane of the same dimensions. For a three-dimensional case, this is illustrated in Fig. 4. The intersection of the elliptic paraboloid with the plane is a curve whose only point of zero derivative is a minimum.

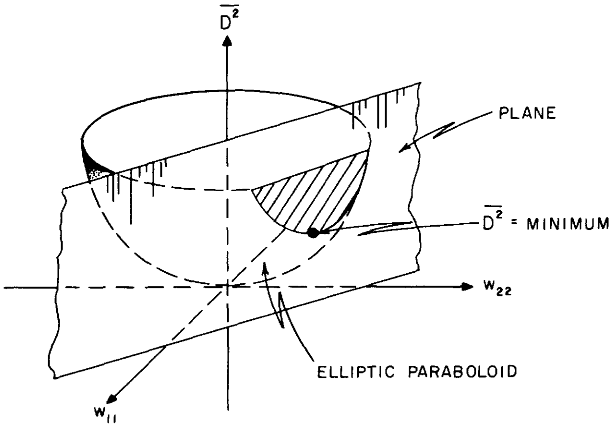


Fig. 4. Geometric interpretation of minimization.

If the variance of a coordinate of the ensemble is large, the corresponding w_{nn} is small, which indicates that small weight is to be given in the over-all measure of distance to a feature of large variation. But if the variance of the magnitude of a given coordinate θ_n is small, its value can be accurately anticipated. Therefore θ_n should be counted heavily in a measure of similarity. It is important to note that in the extreme case, where the variance of the magnitude of a component of the set is zero, the corresponding w_{nn} in Eq. 16 is equal to one, with all other w_{nn} equal to zero. In this case, although Eq. 11 is not a legitimate metric, since it does not satisfy Eq. 2, it is still a meaningful measure of similarity. If any coordinate occurs with identical magnitudes in all members of the set, then it is an "all-important" feature of the set, and nothing else needs to be considered in judging the events similar. Judging membership in a category by such an "all-important" feature may, of course, result in the incorrect inclusion of nonmembers in the category. For instance "red, nearly circular figures" have the color red as a common attribute. The transformation described thus far would pick out "red" as an all-important feature and would judge membership in the category of "red, nearly circular figures" only by the color of the compared object. A red square, for instance, would thus be misclassified and judged to be a "red, nearly circular figure." Given only examples of the category, such results would probably be expected. In Section IV, however, where labeled examples of all categories of interest are assumed given, only those attributes are emphasized in which members of a category are alike and in which they differ from the attributes of other categories.

Note that the weighting coefficients do not necessarily decrease monotonically in the feature weighting, which minimizes the mean-square distance among M given examples of the class. Furthermore, the results of Eqs. 16 or 18 are independent of the particular orthonormal system of coordinates. Equations 16 and 18 simply state that the weighting coefficient is inversely proportional to the variance or to the standard deviation of the ensemble along the corresponding coordinate. The numerical values of the variances, on the other hand, do depend on the coordinate system.

If we use the mathematically more appealing constraint of Eq. 13 in place of that in Eq. 12b, we obtain Eq. 17.

$$\min \overline{D^2} = \min 2 \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 \qquad \prod_{n=1}^{N} w_{nn} = 1 \tag{17a}$$

$$\sum_{n=1}^{N} dw_{nn} \left[w_{nn}\sigma_n^2 - \lambda \prod_{k\neq n}^{N} w_{kk} \right] = 0 \tag{17b}$$

It is readily seen that by applying Eq. 17a, the expression of Eq. 17b is equivalent to Eq. 18a, where the bracketed expression must be zero for all values of n. This substitution leads to Eq. 18b, which may be reduced to Eq. 18c by applying Eq. 17a once more.

$$\sum_{n=1}^{N} dw_{nn} \left(w_{nn}\sigma_n^2 - \frac{\lambda}{w_{nn}} \right) = 0 \tag{18a}$$

$$w_{nn} = \frac{\lambda^{1/2}}{\sigma_n} \tag{18b}$$

$$w_{nn} = \left(\prod_{p=1}^{N} \sigma_p \right)^{1/N} \frac{1}{\sigma_n} \tag{18c}$$

Thus it is seen that the feature weighting coefficient w_{nn} is proportional to the reciprocal standard deviation of the n^{th} coordinates, and thereby lends itself to the same interpretation as before.

2.3 DESCRIBING THE CATEGORY

The set of known members is the best description of the category. Following the practice of probability theory, we can describe this set of similar events by its statistics; the ensemble mean, variance and higher moments can be specified as its characteristic properties. For our purpose a more suitable description of our idea of the category is found in the specific form of the function S of Eq. 1 developed from the set of similar events to measure membership in the category. A marked disadvantage of S is that (in a machine that implements its application) the amount of storage capacity that must be available is proportional to the number of events introduced and is thus a growing quantity. Note that $S(P,\{F_m\})$ can be simplified and expressed in terms of certain statistics of the events, which makes it possible to place an upper bound on the storage requirements demanded of a machine.

Interchanging the order of summations and expanding the squares yield Eq. 19a, which, through the addition and subtraction of $\overline{f_n}^2$, yields Eq. 19b.

$$S(P,\{F_m\}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} W_n^2 (p_n - f_{mn})^2 = \sum_{n=1}^{N} W_n^2 \left[p_n^2 - 2p_n \overline{f_n} + \overline{f_n^2} \right] \tag{19a}$$

$$= \sum_{n=1}^{N} W_n^2 \left[(p_n - \overline{f_n})^2 + \sigma_n^2 \right] = \sum_{n=1}^{N} W_n^2 (p_n - \overline{f_n})^2 + K \tag{19b}$$

We see that the category can be described by first- and second-order statistics of the given samples. This fact also reveals the limitations of the particular class of metrics considered above, for there are classes for which first- and second-order statistics are not sufficient descriptors. It should be pointed out, however, that this is a limitation of the particular, restricted class of metrics just considered, rather than a limitation of the approach used.

By substituting W_n from Eq. 18 in the quadratic form of Eq. 19, we obtain

$$S(P,\{F_m\}) = \lambda \left[\sum_{n=1}^{N} \left(\frac{p_n - \overline{f_n}}{\sigma_n} \right)^2 + N \right] = \left(\prod_{p=1}^{N} \sigma_p \right)^{2/N} \left[\sum_{n=1}^{N} \left(\frac{p_n - \overline{f_n}}{\sigma_n} \right)^2 + N \right] \tag{20}$$

Contours of constant $S(P,\{F_m\})$ are ellipses centered at $\overline{f}$, where $\overline{f}$ is the sample mean, and the diameters of the ellipse are the variances of the samples in the directions of the coordinate axes.

The set of known members of the category appears as the constants $\overline{f_n}$ and σ_n^2, which may be computed once and for all. Since these constants can be updated readily without recourse to the original sample points, the total number of quantities that must be stored is fixed at 2N and is independent of the number of sample points.

2.4 CHOOSING THE OPTIMUM ORTHOGONAL COORDINATE SYSTEM

The labeled events that belong to one category have been assumed given as vectors in an assumed coordinate system that expressed features of the events thought to be relevant to the determination of the category. An optimum set of feature weighting coefficients through which similar events could be judged most similar to one another was then found. It would be purely coincidental, however, if the features represented by the given coordinate system were optimal in expressing the similarities among members of the set. In this section, therefore, we look for a new set of coordinates, spanning the same space and expressing a different set of features that minimize the mean-square distance between members of the set. The problem just stated can be thought of as either enlarging the class of metrics considered thus far in the measure of similarity defined earlier or as enlarging the class of transformations of the space within which class we look for that particular transformation that minimizes the mean-square distance between similar events.

It was proved earlier that the linear transformation that changes the scale of the n^{th} dimension of the space by the factor w_{nn} while keeping the volume of the space constant and minimizing the mean-square distance between the transformed vectors is given by Eq. 22.

$$F' = F[W] \qquad [W] = \begin{bmatrix} w_{11} & & & 0 \\ & w_{22} & & \\ & & \ddots & \\ 0 & & & w_{NN} \end{bmatrix} \tag{22a}$$

and

$$w_{nn} = \left(\prod_{p=1}^{N} \sigma_p \right)^{1/N} \frac{1}{\sigma_n} \tag{22b}$$

The mean-square distance under this transformation is a minimum for the given choice of orthogonal coordinate system. It is given by

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{p=1}^{M} \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nn}^2 (f_{mn} - f_{pn})^2 = \text{minimum} \tag{23}$$

It is possible, however, to rotate the coordinate system until one is found that minimizes this minimum mean-square distance. While the first minimization took place with respect to all choices of the w_{nn} coefficients, we are now interested in further minimizing $\overline{D^2}$ by first rotating the coordinate system so that the above optimum choice of the w_{nn} should result in the absolute minimum distance between vectors. The solution of this search for the optimum transformation can be conveniently stated in the form of the following theorem.

THEOREM

The orthogonal transformation which, after transformation, minimizes the mean-square distance between a set of vectors, subject to the constraint that the volume of the space is invariant under transformation, is a rotation [C] followed by a diagonal transformation [W]. The rows of the matrix [C] are eigenvectors of the covariance matrix [U] of the set of vectors, and the elements of [W] are those given in Eq. 22b, where σ_p is the standard deviation of the coefficients of the set of vectors in the direction of the p^{th} eigenvector of [U].

PROOF

Expanding the square of Eq. 23 and substituting the values of w_{nn} result in Eq. 24, which is to be minimized over all choices of the coordinate system.

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sum_{p=1}^{M} \sum_{m=1}^{M} \left(f_{mn}^2 + f_{pn}^2 - 2f_{mn}f_{pn}\right) \tag{24a}$$

$$= \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left(\overline{f_n^2} - \overline{f_n}^2\right) = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 \tag{24b}$$

$$= \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sigma_p^2\right]^{1/N} \tag{24c}$$

Let the given coordinate system be transformed by the matrix [C]:

$$[C] = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix} \quad \text{where} \sum_{n=1}^{N} c_{pn}^2 = 1 \qquad p = 1, 2, \ldots, N \tag{25}$$

Equation 24 is minimized if the bracketed expression in Eq. 24c, which we shall name β, is minimized.

$$\beta = \prod_{p=1}^{N} \sigma_p^2 = \prod_{p=1}^{N} \left[\frac{1}{M} \sum_{m=1}^{M} (f'_{mp})^2 - \left(\frac{1}{M} \sum_{m=1}^{M} f'_{mp}\right)^2\right] \tag{26a}$$

where

$$f'_{mp} = \sum_{n=1}^{N} f_{mn} c_{pn} \tag{26b}$$

Substituting Eq. 26b into Eq. 26a, we obtain

$$\beta = \prod_{p=1}^{N} \left[\sum_{n=1}^{N} \sum_{s=1}^{N} \frac{1}{M} \sum_{m=1}^{M} f_{mn} f_{ms} c_{pn} c_{ps} - \left(\sum_{n=1}^{N} \overline{f_n}\, c_{pn}\right)^2\right] \tag{27}$$

in which the averaging is understood to be over the set of M vectors. The squared expression may be written as a double sum and the entire equation simplified to

$$\beta = \prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} \left(\overline{f_n f_s} - \overline{f_n}\,\overline{f_s} \right) c_{pn} c_{ps} \tag{28}$$

But $\left(\overline{f_n f_s} - \overline{f_n}\,\overline{f_s} \right) = u_{ns} = u_{sn}$ is an element of the covariance matrix [U]. Hence we have

$$\beta = \prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \tag{29}$$

Using the method of Lagrange multipliers to minimize β in Eq. 29, subject to the constraint of Eq. 25, we obtain Eq. 30 as the total differential of β. The differential of the constraint, γ, is given in Eq. 31.

$$d\beta(c_{11}c_{12}\cdots c_{NN})$$
$$= \sum_{\ell=1}^{N} \sum_{q=1}^{N} \left[\prod_{p\neq\ell}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right] \frac{\partial}{\partial c_{\ell q}} \left(\sum_{a=1}^{N} \sum_{b=1}^{N} u_{ab} c_{\ell a} c_{\ell b} \right) dc_{\ell q} \tag{30}$$

$$d\gamma_\ell = 2 \sum_{q=1}^{N} c_{\ell q} dc_{\ell q} \qquad \ell = 1, 2, \ldots, N \tag{31}$$

By way of an explanation of Eq. 30, note that when Eq. 29 is differentiated with respect to $c_{\ell q}$, then all of the factors in the product in Eq. 29, where $p \neq \ell$, are simply constants. Carrying out the differentiation stated in Eq. 30, we obtain

$$d\beta = 2 \sum_{\ell=1}^{N} \sum_{q=1}^{N} dc_{\ell q} \left[\sum_{b=1}^{N} c_{\ell b} u_{qb} \right] \prod_{p\neq\ell}^{N} \left[\sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right] \tag{32}$$

Now let

$$\prod_{p\neq\ell}^{N} \cdot \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} = A_\ell \tag{33}$$

and note that since $p \neq \ell$, A_ℓ is just a constant as regards optimization of any $c_{\ell x}$.

In accordance with the method of Lagrange multipliers, each of the N constraints of Eq. 31 is multiplied by a different arbitrary constant B_ℓ and is added to dβ as shown below.

$$d\beta + \sum_{\ell=1}^{N} B_\ell d\gamma_\ell = 0 = 2 \sum_{\ell=1}^{N} \sum_{q=1}^{N} dc_{\ell q} \left[\left(\sum_{b=1}^{N} c_{\ell b} u_{qb} \right) A_\ell + B_\ell c_{\ell q} \right] \tag{34}$$

By letting $-\lambda_\ell = B_\ell / A_\ell$ and by recognizing that $dc_{\ell q}$ is arbitrary, we get

$$\sum_{b=1}^{N} c_{\ell b} u_{qb} - \lambda_\ell c_{\ell q} = 0 \qquad q=1,2,\ldots,N;\ \ell=1,2,\ldots,N \tag{35}$$

Let the ℓ^{th} row of the [C] matrix be the vector C_ℓ. Then Eq. 35 can be written as the eigenvalue problem of Eq. 36 by recalling that $u_{qb} = u_{bq}$.

$$C_\ell [U - \lambda_\ell I] = 0 \qquad \ell = 1,2,\ldots,N \tag{36}$$

Solutions of Eq. 36 exist only for N specific values of λ_ℓ. The vector C_ℓ is an eigenvector of the covariance matrix [U]. The eigenvalues λ_ℓ are positive, and the corresponding eigenvectors are orthogonal, since the matrix [U] is positive definite. Since the transformation [C] is to be nonsingular, the different rows C_ℓ must correspond to different eigenvalues of [U]. It may be shown that the only extremum of β is a minimum, subject to the constraint of Eq. 25. Thus the optimum linear transformation that minimizes the mean-square distance of a set of vectors while keeping the volume of the space constant is given by Eq. 37, where rows of [C] are eigenvectors of the covariance matrix [U].

$$\begin{bmatrix} a_{11} a_{12} & \cdots & a_{1N} \\ a_{21} a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & \cdots \\ a_{N1} a_{N2} & \cdots & a_{NN} \end{bmatrix} = \begin{bmatrix} c_{11} c_{12} & \cdots & c_{1N} \\ c_{21} c_{22} & \cdots & c_{2N} \\ \cdots & \cdots & \cdots \\ c_{N1} c_{N2} & \cdots & c_{NN} \end{bmatrix}^T \begin{bmatrix} w_{11} & & \\ & w_{22} & \\ & & \ddots \\ & & & w_{NN} \end{bmatrix}$$

$$= \begin{bmatrix} w_{11}c_{11} & w_{22}c_{21} & \cdots & w_{NN}c_{N1} \\ w_{11}c_{12} & w_{22}c_{22} & \cdots & w_{NN}c_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ w_{11}c_{1N} & w_{22}c_{2N} & \cdots & w_{NN}c_{NN} \end{bmatrix} \tag{37}$$

The numerical value of the minimum mean-square distance may now be computed as follows. The quantity $\overline{D^2}$ was given in Eq. 24c, which is reproduced here as Eq. 38:

$$\overline{D^2} = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sigma_p^2 \right]^{1/N} = \frac{M}{(M-1)} 2N(\beta)^{1/N} \tag{38}$$

Substituting β from Eq. 29, we obtain

$$\overline{D^2} = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right]^{1/N} \tag{39}$$

But from Eq. 35 we see that min $\overline{D^2}$ may be written as

$$\min \overline{D^2} = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sum_{n=1}^{N} \lambda_p c_{pn}^2\right]^{1/N} = \frac{M}{(M-1)} 2N \left(\prod_{p=1}^{N} \lambda_p\right)^{1/N} \tag{40}$$

in which the constraint of Eq. 25 has been used.

It should be noted that the constraint of Eq. 25 is not, in general, a constant volume constraint. Instead, the constraint holds the product of the squared lengths of the sides of all N-dimensional parallelepipeds a constant. If, as in the solution just obtained, the transformation [C] is orthogonal, the volume is maintained constant. A subset of the constant volume transformations, T_v (Fig. 5), are the orthogonal transformations T_o of constant volume of which the optimum was desired. The solution presented here found the optimum transformation among a set of T_L that contains orthogonal transformations of constant volume but is not necessarily constant volume for those that are nonorthogonal. The solution given here, therefore, is optimum among the constant volume transformations $T_v \cap T_L$ shown as the shaded area in Fig. 5. This intersection is a larger set of transformations than that for which the optimum was sought.

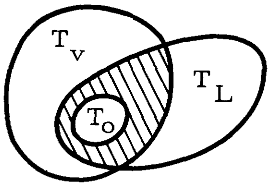


Fig. 5. Sets of transformations.

The methods of this section are optimal in measuring membership in categories of certain types. Suppose, for instance, that categories are random processes that generate members with multivariate Gaussian probability distributions of unknown means and variances. In Appendix B we show that the metric developed here measures contours of equal a posteriori probabilities. Given the set of labeled events, the metric specifies the locus of points that are members of the category in question with equal probability.

2.5 SUMMARY

Categorization, the basic problem of pattern recognition, is regarded as the process of learning how to partition the signal space into regions where each contains points of only one category. The notion of similarity between a point and a set of points of a category plays a dominant role in the partitioning of signal space. Similarity of a point to a set of points is regarded as the average "distance" between the point and the set. The sense in which distance is understood is not specified, but the optimum sense is thought to be that which (by the optimum method of measuring distance) clusters most highly those points that belong to the same category. The mean-square distance between points of a category is a measure of clustering. An equivalent alternate interpretation of similarity (not as general as the interpretation above) is that the transformation that optimally clusters like points, subject to suitable criteria to ensure the nontriviality of the transformations, is instrumental in exhibiting the similarities between points of a set. In particular, the optimum orthogonal transformation, and hence a non-Euclidean method of measuring distance, is found that minimizes the mean-square distance between a set of points, if the volume of the space is held constant to ensure nontriviality. The resulting measure of similarity between a point P and a set $\{F_m\}$ is

$$S(P,\{F_m\}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \left[\sum_{s=1}^{N} a_{ns}(p_s - f_{ms}) \right]^2 \tag{41}$$

where a_{ns} is given in the theorem of section 2.4.

Classification of an arbitrary point P into one of two categories, F or G, is accomplished by the decision rule given in Eq. 42, where the functions S_f and S_g are obtained from samples of F and samples of G, respectively.

$$\begin{aligned} &\text{decide } P \in F \text{ if } S_f(P,\{F_m\}) < S_g(P,\{G_m\}) \\ &\text{decide } P \in G \text{ if } S_f(P,\{F_m\}) > S_g(P,\{G_m\}) \end{aligned} \tag{42}$$

III. CATEGORIZATION

3.1 THE PROCESS OF CLASSIFICATION

Pattern recognition consists of the twofold task of "learning" what the category or class is to which a set of events belongs and of deciding whether or not a new event belongs to the category. In this section, details of the method of accomplishing these two parts of the task are discussed, subject to the limitations on recognizable categories imposed by the assumptions stated earlier. These details are limited to the application of the special method of Section II.

In the following section two distinct modes of operation of the recognition system will be distinguished. The first consists of the sequential introduction of a set of events, each labeled according to the category to which it belongs. During this period, we want to identify the common pattern of the inputs that allows their classification into their respective categories. As part of the process of learning to categorize, the estimate of what the category is must also be updated to include each new event as it is introduced. The process of updating the estimate of the common pattern consists of recomputing the new measures of similarity so that they will include the new, labeled event on which the quantitative measures of similarity are based.

During the second mode of operation the event P, which is to be classified, is compared to each of the sets of labeled events by the measure of similarity found best for each set. The event is then classified as a member of that category to which it is most similar.

It is not possible to state with certainty that the pattern has been successfully learned or recognized from a set of its samples because information is not available on how samples were selected to represent the class. Nevertheless, it is possible to obtain a quantitative indication of our certainty of having obtained a correct method of determining membership in the category from the ensemble of similar events. As each new event is introduced, its similarity to the members of the sets already presented is measured by the function S defined in Section II. The magnitude of the number S indicates how close the new event is to those already introduced. As S is refined and, with each new example, improves its ability to recognize the class, the numerical measure of similarity between new examples and the class will tend to decrease, on the average. Strictly speaking, of course, this last statement cannot be true, in general. It may be true only if the categories to be distinguished are separable by functions S taken from the class that we have considered; even under this condition the statement is true only if certain assumptions are made regarding the statistical distribution of the samples on which we learn.* In cases in which no knowledge regarding the satisfaction of either of these two requirements exists, the convergence of the similarity as the sample size is

*It is true for Gaussian processes.

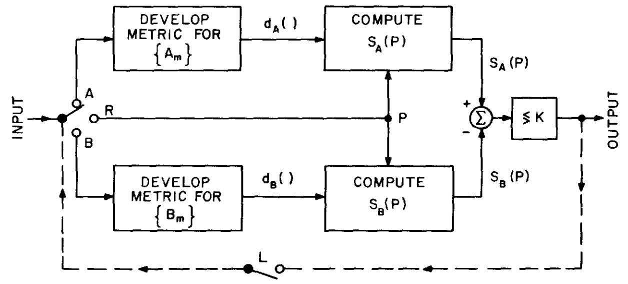


Fig. 6. Elementary block diagram of the classification process.

increased is simply wishful thinking the heuristic justification of which is based on the minimization problem solved in developing S.

Figure 6 illustrates the mechanization of the learning and recognition modes of the special classificatory process discussed thus far. For the sake of clarity, the elementary block diagram of the process distinguishes only between two categories of events, but it can be readily extended to distinguish between an arbitrary number of categories. It should be noted that one of the categories may be the complement of all others. The admission of such a category into the set is only one of the ways in which a machine that is always forced to classify events into known categories may be made to decide that an event does not belong to any of the desired ones; it belongs to the category of "everything else." Samples of "everything else" must, of course, be given.

During the first mode of operation, the input to the machine is a set of labeled events. Let us follow its behavior through an example. Suppose that a number of events, some belonging to set A and some to set B, have already been introduced. According to the method described in Section II, therefore, the optimum metrics (one for each class) that minimize the mean-square distance between events of the same set have been found. As a new labeled event is introduced (say that it belongs to set A), the switch at the input is first turned to the recognition mode R so that the new event P can be compared to set A as well as to set B through the functions $S_A(P) = S(P, \{A_m\})$ and $S_B(P)$, which were computed before the introduction of P. The comparison of S_A and S_B with a threshold K indicates whether the point P would be classified as belonging to A or to B from knowledge available up to the present. Since the true membership of P is known (the event is labeled), we can now determine whether P would be classified correctly or incorrectly. The input switch is then turned to A so that P, which indeed belongs to A, can be included in the computation of the best metric of set A.

When the next labeled event is introduced (let us say that it belongs to set B), the input switch is again turned to R in order to test the ability of the machine to classify the new event correctly. After the test, the switch is turned to B so that the event can be included among examples of set B and the optimum function S_B can be recomputed. This procedure is repeated for each new event, and a record is kept of the rate at which incorrect classifications would be made on the known events. When the training period is completed, presumably as a result of satisfactory performance on the selection of known events (sufficiently low error rate), the input switch is left in the recognition mode.

3.2 LEARNING

"Supervised learning" takes place in the interval of time in which examples of the categories generate ensembles of points from which the defining features of the classes are obtained by methods previously discussed. "Supervision" is provided by an outside source such as a human being who elects to teach the recognition of patterns by examples and who selects the examples on which to learn.

"Unsupervised learning," by contrast, is a method of learning without the aid of such an outside source. It is clear, at least intuitively, that the unsupervised learning of membership in specific classes cannot succeed unless it is preceded by a period of supervision during which some concepts regarding the characteristics of classes are established. A specified degree of certainty concerning the patterns has been achieved in the form of a sufficiently low rate of misclassification during the supervised learning period. The achievement of the low misclassification rate, in fact, can be used to signify the end of the learning period, after which the system that performs the operations indicated in Fig. 6 can be left to its own devices. It is only after this supervised interval of time that the system can be usefully employed to recognize, without outside aid, events as belonging to one or another of the categories.

Throughout the period of learning on examples, each example is included in its proper set of similar events that influence the changes of the measures of similarity. After supervised activity has ceased, events introduced for classification may belong to any of the categories; and no outside source informs the machine of the correct category. The machine itself, operating on each new event, however, can determine, with the already qualitatively specified probability of error, to which class the event should belong. If the new event is included in the set exemplifying this class, the function measuring membership in the category has been altered. Unsupervised learning results from the successive alterations of the metrics, brought about by the inclusion of events into the sets of labeled events according to determinations of class membership rendered by the machine itself. This learning process is instrumented by the dotted line in Fig. 6, which, when the learning switch L is closed, allows the machine's decisions to control routing of the input to the various sets.

To facilitate the illustration of some implications of the process described above, consider the case in which recognition of membership in a single class is desired and all labeled events are members of only that class. In this case, classification of events as members or nonmembers of the category degenerates into the comparison of the similarity S with a threshold T. If S is greater than T, the event is a nonmember; if S is less than T, the event is said to be a member of the class. Since the machine decides that all points of the signal space for which S is less than T are members of

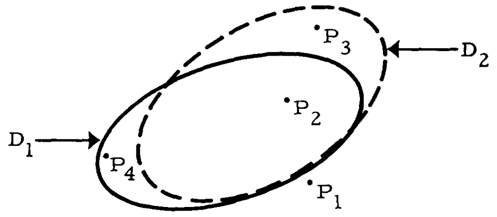


Fig. 7. Unsupervised learning.

the class, the class – as far as the machine is concerned – is the collection of points that lie in a given region in the signal space. For the specific function S of Section II, this region is an ellipsoid in the N-dimensional space.

Unsupervised learning is illustrated graphically in Fig. 7. The two-dimensional ellipse drawn with a solid line signifies the domain D_1 of the signal space in which any point yields $S < T$. This domain was obtained during supervised activity. If a point P_1 is introduced after supervised learning, so that P_1 lies outside D_1, then P_1 is merely rejected as a nonmember of the class. If point P_2 contained in D_1 is introduced, however, it is judged a member of the class and is included in the set of examples used to generate a new function S and a new domain D_2, designated by the dotted line in Fig. 7. A third point P_3, which was a nonmember before the introduction of P_2, becomes recognized as a member of the class after the inclusion of P_2 in the set of similar events.

Although the tendency of this process of "learning" is to perpetuate the original domain, it has interesting properties worth investigating. The investigation of unsupervised learning would form the basis for a valuable continuation of the work presented herein.

Before leaving the subject of unsupervised learning, we point out that as the new domain D_2 is formed, points such as P_4 in Fig. 7 become excluded from the class. Such an exclusion from the class is analogous to "forgetting" because of lack of repetition. Forgetting is the characteristic of not recognizing P_4 as a member of the class, although at one time it was recognized as belonging to it.

3.3 THRESHOLD SETTING

In the classification of an event P, the mean-square distance between P and members of each of the categories is computed. The distance between P and members of a category C is what we called "similarity", $S_C(P)$, in which the "sense" in which "distance" is understood depends on the particular category in question. We then stated that, in a manner analogous to decisions based on maximum likelihood ratios, the point P is classified as a member of the category to which it is most similar. Hence, P belongs to category C if $S_C(P)$ is less than $S_X(P)$, where X is any of the other categories.

Since in this special theory the function $S_C(P)$, which measures membership in category C, was developed by maximally clustering points of C without separating them from points of other sets, there is no guarantee, in general, that a point of another set B may

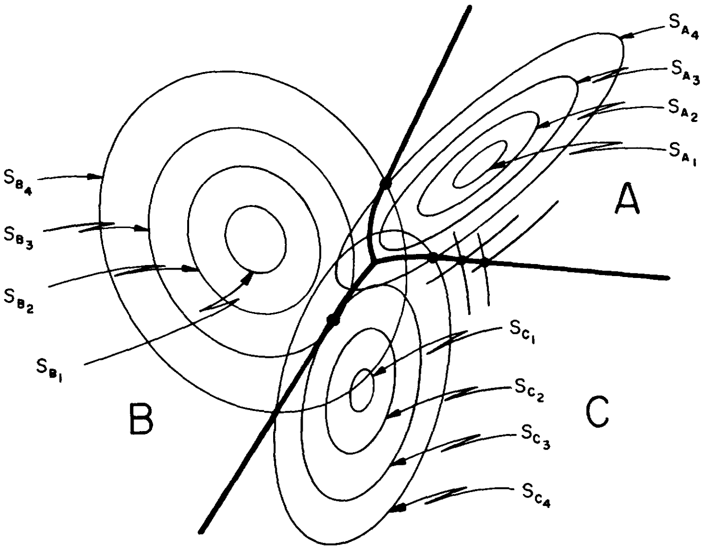


Fig. 8. Categorization.

not be closer to C than to B. This is guaranteed only if points of the sets occupy disjointed regions. A graphical illustration that clarifies the comparison of similarities of a point to the different categories is shown in Fig. 8. In this figure the elliptical contours $S_{A_1}(P)$, $S_{A_2}(P)$, ..., indicate the locus of points P in the signal space that is at a mean-square distance of 1, 2, ..., from members of category A. The loci of these points are concentric ellipsoids in the N-dimensional signal space, shown here in only two dimensions. Similarly, $S_{B_1}(P)$, $S_{B_2}(P)$, ..., and $S_{C_1}(P)$, $S_{C_2}(P)$, ..., are the loci of those points whose mean-square distance from categories B and C, respectively, are 1, 2, ... Note carefully that the sense in which distance is measured at each of the categories differs as is indicated by the different orientations and eccentricities of the ellipses. The heavy line shows the locus of points that are at equal mean-square distances from two or more sets according to the manner in which distance is measured to each set. This line, therefore, defines the boundary of each of the categories.

At this point in the discussion it will be helpful to digress from the subject of thresholds and dispel some misconceptions that Fig. 8 might create regarding the general nature of the categories found by the method that we described. Recall that one of the possible ways in which a point not belonging to either category could be so classified was by establishing a separate category for "everything else" and assigning the point to the category to which its mean-square distance is smallest. Another, perhaps more practical, method is to call a point a member of neither category if its mean-square distance to the set of points of any class exceeds some threshold value. If this threshold value is set, for example, at a mean-square distance of 3 for all of the categories in Fig. 8, then points belonging to A, B, and C will lie inside the three ellipses shown in Fig. 9.

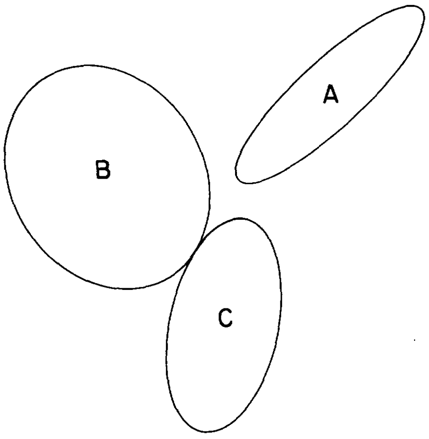


Fig. 9. Categorization with threshold.

It is readily seen, of course, that there is no particular reason why one given minimum mean-square distance should be selected instead of another; or, for that matter, why this minimum distance should be the same for all categories. Many logical and useful criteria may be selected for determining the optimum threshold setting. Here, only one criterion will be singled out as particularly useful. This criterion requires that the minimum thresholds be set so that most of the labeled points fall into the correct category. This is a fundamental criterion, for it requires the system to work best by making the largest number of correct decisions. In decision theory the threshold value depends on the a priori probabilities of the categories and on the costs of false alarm and false dismissal.

The criterion of selecting a threshold that will make the most correct classifications can be applied to our earlier discussions in which the boundary between categories was determined by equating the similarities of a point to two or more categories. In the particular example of Fig. 6, where a point could be a member of only one of two categories A and B, the difference $S_A - S_B = 0$ formed the dividing line. There is nothing magical about the threshold zero; we might require that the dividing line between the two categories be $S_A - S_B = K$, where K is a constant chosen from other considerations. A similar problem in communication theory is the choice of a signal-to-noise ratio that serves as the dividing line between calling the received waveform "signal" or calling it "noise." It is understood, of course, that signal-to-noise ratio is an appropriate criterion on which to base decisions (at least in some cases), but the particular value of the ratio to be used as a threshold level must be determined from additional requirements. In communication theory these are usually requirements on the false-alarm or false-dismissal rates. In the problem of choosing the constant K, we may require that it be selected so that most of the labeled points lie in the correct category.

3.4 PRACTICAL CONSIDERATIONS

In considering the instrumentation of the process of categorization previously described, two main objectives of the machine design must receive careful consideration. The first is the practical requirement that all computations involved in either the learning or the recognition mode of the machine's operation shall be performed as rapidly as possible. It is especially desirable that the classification or recognition of a new event be implemented in essentially real time. The importance of this requirement is readily appreciated if the classificatory technique is considered in terms of an application such as the automatic recognition of speech events, an important part of voice-controlled phonetic typewriters. The second major objective, not unrelated to the first, is that the storage capacity required of the machine have an upper bound, thus assuring that the machine is of finite and predetermined size. At first glance it seems that the instrumentation of the machine of Fig. 6 requires a storage capacity proportional to the number of events encountered during the machine's experience. This seems so because the set of labeled events on which the computations are carried out must be stored in the machine. We shall now show, however, that all computations can be performed from knowledge of only certain statistics of the set of labeled events, and that these statistics can be recomputed to include a new event without knowledge of the original set. Therefore, it is necessary to store only these statistics, the number of which is independent of the number of points in the set.

It will be recalled that there are two instances when knowledge of the data matrix is necessary. The data matrix [F], given in Eq. 43, is the M × N matrix of coefficients that results when the M given examples of the same category are represented as N-dimensional vectors.

$$[F] = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ f_{M1} & f_{M2} & \cdots & f_{MN} \end{bmatrix} \tag{43}$$

The first use of this matrix occurs in the computation of the optimum orthogonal transformation or metric that minimizes the mean-square distance of the set of like events. This transformation is stated in the theorem in Section 2. 4 and is given in Eq. 37 as the product of an orthonormal and a diagonal transformation. Rows of the orthonormal transformation [C] are eigenvectors of the covariance matrix [U] computed from the data matrix of Eq. 43, and elements of the diagonal matrix [W] are the reciprocal standard deviations of the data matrix after it has been transformed by the orthonormal transformation [C].

The second use of the matrix [F] occurs when an unclassified event P is compared to the set by measuring the mean-square distance between P and points of the set after both the point and the set have been transformed. This comparison is replaced by the measurement of the distance between the transformed point P and the mean vector of the set after transformation. The quantities of interest in this computation are the mean, the mean-square, and the standard deviation of the elements in the columns of the data matrix after the orthonormal transformation.

Reduction of the necessary storage facility of the machine can be accomplished if only the covariance matrix, the means, the mean squares, and the standard deviations of the transformed data matrix are used in the computations, and if these can be recomputed without reference to the original data matrix. The expression of the above quantities when based on M + 1 events may be computed from the corresponding quantity based on M events and a complete knowledge of the M + 1st event itself. Let us look at the method of computation.

(a) The covariance matrix of M + 1 events

The general coefficient of the covariance matrix $[U]$ of the set of events given by the data matrix $[F]$ is given by

$$u_{ns} = u_{sn} = \overline{f_n f_s} - \overline{f_n}\,\overline{f_s} \tag{44}$$

Note, incidentally, that the matrix $[U]$ may be written as in Eq. 45a, where the matrix $[J]$ has been introduced for convenience. As a check, let us compute the general element u_{ns}.

$$[U] = \frac{1}{M}[F-J]^T[F-J] \tag{45a}$$

in which

$$[J] = \begin{bmatrix} \overline{f_1} & \overline{f_2} & \cdots & \overline{f_N} \\ \overline{f_1} & \overline{f_2} & \cdots & \overline{f_N} \\ \cdots & \cdots & \cdots & \cdots \\ \overline{f_1} & \overline{f_2} & \cdots & \overline{f_n} \end{bmatrix} \tag{45b}$$

The nth column of the $[F-J]$ matrix, which becomes the nth row of its transposition is given in Eq. 46, as well as the sth column of $[F-J]$. The product is the covariance matrix coefficient u_{ns}.

$$u_{ns} = \frac{1}{M}\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ f_{1n}-\overline{f_n}, & f_{2n}-\overline{f_2}, & \ldots, & f_{Mn}-\overline{f_n} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}\begin{bmatrix} \cdot & f_{1s}-\overline{f_s} & \cdot \\ \cdot & f_{2s}-\overline{f_s} & \cdot \\ \cdot & \vdots & \cdot \\ \cdot & f_{Ms}-\overline{f_s} & \cdot \end{bmatrix} \tag{46}$$

$$u_{ns} = \frac{1}{M}\sum_{m=1}^{M}(f_{mn}-\overline{f_n})(f_{ms}-\overline{f_s}) = \overline{f_n f_s} - \overline{f_n}\,\overline{f_s} \tag{47}$$

Now to compute the covariance based on M + 1 events, $u_{ns}(M+1)$, it is convenient to store the N means $\overline{f_n}$ for all values of n. It is also convenient to store the N(N+1)/2 independent values of $\overline{f_n f_s}$. Both of these quantities may be updated readily as a new event is introduced. The mean $\overline{f_n}^{M+1}$, based on M + 1 events may be obtained from the mean based on only M events, $\overline{f_n}^M$, from Eq. 48a, and $\overline{f_n f_s}^{M+1}$ may be obtained from Eq. 48b.

$$\overline{f_n}^{M+1} = \frac{M\overline{f_n}^M + f_{M+1,n}}{M+1} \tag{48a}$$

$$\overline{f_n f_s}^{M+1} = \frac{M\overline{f_n f_s}^M + f_{M+1,n} f_{M+1,s}}{M+1} \tag{48b}$$

Here, the superscript of the ensemble average indicates the number of events partaking in the averaging, and $f_{M+1,n}$ is the n^{th} coefficient of the M + 1st event. We now have everything necessary for computing the new covariance coefficients. The storage facility required thus far is $N(N+3)/2 + 1$. The +1 is used for storing the number M. If the covariance matrix is also stored, the necessary number of storage locations is $(N+1)^2$; this makes use of the fact that both [U] and $[F^TF]$ are symmetric matrices.

From the matrix [U] the orthonormal transformation [C] may be found by solving the eigenvalue problem $[C][U-\lambda I] = 0$. The matrix [C] has to be stored, requiring an additional N^2 storage locations.

(b) Mean of the p^{th} column of the [F][C] = [F'] matrix

As stated earlier, one of the quantities of interest is the mean of the elements in a column of the data matrix after its orthonormal transformation with [C]. The general element of the [F'] matrix is f'_{mp} given in Eq. 26b and in Eq. 49a, and its mean is given in Eq. 49b.

$$f'_{mp} = \sum_{n=1}^{N} f_{mn} c_{pn} \tag{49a}$$

$$\overline{f'_p} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} f_{mn} c_{pn} = \sum_{n=1}^{N} \overline{f_n} c_{pn} \tag{49b}$$

No additional storage is required for computing $\overline{f'_p}$ because all of the factors of Eq. 49b are already known. An additional N locations must be made available to store the N means, however.

(c) Mean square of p^{th} column of the [F'] matrix

The mean-square value of elements of the p^{th} column of [F'] is given in Eqs. 50a and 50b.

$$\overline{f'^2_p} = \frac{1}{M}\sum_{m=1}^{M} f'^2_{mp} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{s=1}^{N} f_{mn} f_{ms} c_{pn} c_{ps} \tag{50a}$$

$$\overline{f'^2_p} = \sum_{n=1}^{N}\sum_{s=1}^{N} \overline{f_n f_s} c_{pn} c_{ps} \tag{50b}$$

No additional storage is necessary for this computation. An additional N locations, however, must be available to store $\overline{f'^2_p}$.

(d) Standard deviation of the p^{th} column of the [F'] matrix

The only remaining quantity necessary in the instrumentation of the recognition system is the reciprocal standard deviation of the p^{th} column of [F'], as stated in the theorem of Section II. The standard deviation and the elements of the diagonal matrix [W] are given by Eq. 51, where all the quantities are already known. An additional N locations are needed to store their values, however.

$$w_{pp} \propto \frac{1}{\sigma'_p} = \frac{1}{\left(\overline{f'^2_p} - \overline{f'_p}^2\right)^{1/2}} \tag{51}$$

The total number of storage locations is about $2N^2$ for each of the categories to which events may belong. If the number of examples M of a category is less than the number of dimensions N of the space in which they are represented, the required number of storage locations is about $2M^2$. In order to utilize this further reduction of storage and computational time, however, the M events must be re-expressed in a new coordinate system obtained through the Schmidt orthogonalization of the set of M vectors representing the examples of the set. In the beginning of the learning process, when the number of labeled events is very much smaller than the number of dimensions of the space, the saving achieved by Schmidt orthogonalization is very significant.

A practical remark worthy of mention is that at the beginning of the learning process, when M is less than N, the solution of the eigenvalue problem $[U-\lambda I] = 0$ may be greatly simplified by recognition of the fact that [U] is singular if $M < N$. The nonzero eigenvalues of [U] in Eq. 45a are identical with the eigenvalues of the matrix $[F-J][F-J]^T$:

$$\text{Nonzero eigenvalues of } [F-J]^T[F-J] = \text{eigenvalues of } [F-J][F-J]^T \tag{52}$$

The first of the matrices is an N × N, while the second is an M × M, matrix. There are N-M zero eigenvalues of the first matrix; the computational advantage of working with the second matrix for $M < N$ is therefore significant.

Let us look at the nature of the solution obtained with the two constraints of Eqs. 12b and 13. It should be noted, first of all, that if the number of points in a set is equal to or less than the number of dimensions in which they are expressed, then a hyperplane of one less dimension can always be passed through the points. Along any direction orthogonal to this hyperplane, the projections of points of the set F are equal. Along such a direction, therefore, the variance of the given points is zero, leading to a zero eigenvalue of the covariance matrix. This results in calling the corresponding eigenvector (the direction about which the variance is zero) an "all-important" feature. The feature weighting coefficient W_n is thus unity or infinity, depending on which of the above two constraints was applied. If the second, or constant volume, constraint were used, each point of the set F used in learning would be correctly identified, and its distance to the set F would be zero by the optimum metric. At the same time, the metric classifies each point of another category G as a nonmember of F. A new member of category F, on the other hand, would probably be misclassified, since it is unlikely that the new member of F would have exactly the same projection along the eigenvector as the other members had displayed. This misclassification would not occur if the number of examples of the category F exceeded the number of dimensions in which they were expressed. There are several methods to prevent misclassification; for example, if the first constraint were applied, misclassification of members of F would not occur.

Another fact of some importance that should be brought to the reader's attention is the physical significance of the eigenvectors. The vector with the smallest eigenvalue or largest feature weighting coefficient designates that feature of members of the set in which the members are most similar. This is not equivalent to the feature that is most similar to members of the set. The former is a solution of a problem in which we wish to find a direction along which the projections of the set, on the average, are most nearly the same. The second is a solution of a problem in which we wish to find the direction along which the projections of the set are largest, on the average. The desired direction, in the first case, is the eigenvector of the covariance matrix with the smallest eigenvalue; in the second case, it is the eigenvector of the correlation matrix $[F^T F]$ with the largest eigenvalue. It can be shown that the latter problem is equivalent to finding the set of orthonormal functions in which a process is to be expanded so that the truncation error, which results when only a finite number of terms of the expansion are retained, should be minimized, on the average. The set of functions having this property are eigenfunctions of the correlation function of the process, and they are arranged in the order of decreasing eigenvalues.

The important concepts of this section will now be summarized. Pattern recognition consists of the twofold task of "learning" what the category is to which a set of events belongs; and of deciding whether or not a new event belongs to the category. "Learning", for the simple situation in which similarity to a class of things is determined solely from examples of the class, may be instrumented in the form of the diagram of Fig. 6. In this diagram, "learning" consists of the construction of metrics or the development of linear transformations that maximize the clustering of points that represent similar events. A distinction is made between "supervised learning" (learning on known examples of the class) and "unsupervised learning" (learning through use of the machine's own experience). The convergence of a learning process to correct category recognition, in most cases, probably cannot be guaranteed. The problem of threshold setting for partitioning the signal space is likened to the similar problem in the detection of noisy signals, and may be solved as an extremum problem. Finally, some practical considerations of importance in the mechanization of the decision process are discussed. It is shown that only finite storage capacity is required of the machine that instruments the techniques, and that the amount of storage has an upper bound that depends on the number of dimensions of the signal space.

IV. CATEGORIZATION BY SEPARATION OF CLASSES

4.1 OPTIMIZATION CRITERIA

The central concept of the special theory of similarity described in the preceding sections is that nonidentical events of a common category may be considered close by some method of measuring distance. This measure of distance is placed in evidence by that transformation of the signal space that brings together like events by clustering them most. In this special theory no attempt was made to ensure that the transformations that were developed should separate events of different categories.

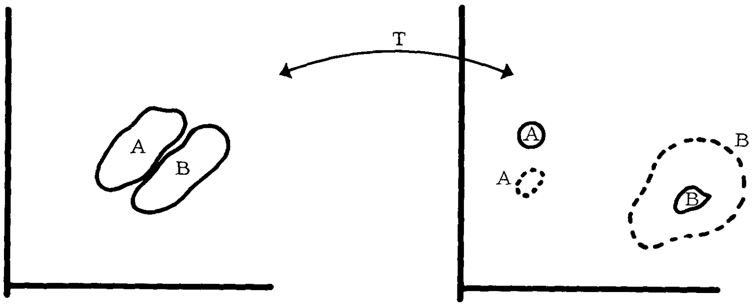


Fig. 10. Separation of classes.

We shall now introduce criteria for developing optimum transformations that not only cluster events of the same class but also separate those that belong to different classes. Consider, for example, the transformation that maximizes the mean-square distance between points that belong to different classes while it minimizes the mean-square distance between points of the same class. The effect of such a transformation is illustrated in Fig. 10, where like events have been clustered through minimization of intraset distances, and clusters have been separated from each other through the maximization of interset distances. The transformation that accomplishes the stated objectives can be specified by the following problems.

Problem 1

Find the transformation T within a specified class of transformations that maximizes the mean-square interset distance subject to the constraint that the sum of the mean-square interset and intraset distances is held constant.

Note that for the sake of simplifying the mathematics, the minimization of intraset distances was converted to a constraint on the maximization problem. If interset distances are maximized, and the sum of interset and intraset distances is constant, then it follows that intraset distances are minimized. We may impose the additional constraint that the mean-square intraset distance of each class is equal, thereby avoiding the possible preferential treatment of one class over another. Without the latter constraint the situation indicated with dotted lines in Fig. 10 may occur when minimization of the sum of intraset distances may leave one set more clustered than the other.

This criterion of optimization is given as an illustrative example of how one may convert the desirable objective of separation of classes to a mathematically expressible and solvable problem. Several alternate ways of stating the desired objectives as well as of choosing the constraints are possible. For example, the mean-square intraset distance could be minimized while holding the interset distances constant. Another alternative is to minimize intraset distances while holding the distances between the means a constant. It can be shown that the solution of this minimization problem results in a transformation which, together with the decision rule postulated to differentiate between members of the different classes, is a sound result and has a counterpart in decision theory.

The optimization criterion just discussed suggests a different block diagram for the process of categorization from that shown in Fig. 6. Here only a single transformation is developed, which results in only a single metric with which to measure distance to all of the classes. The classification of an event P is accomplished, as before, by noting to which of the classes the event is most similar. The only difference is that now, similarity to each class is measured in the same sense, in the sense exhibited by the transformation that maximally separated events of different categories, on the average.

Problem 2

A second, even more interesting criterion for optimum categorization is the optimization of the classificatory decision on the labeled events. Classificatory decisions are ultimately based on comparing the similarity S (mean-square distance) of the event P with the known events of each class. If P is chosen as any member of Class A, for example, we would like to have $S|P,\{A_m\}| < S|P,\{B_m\}|$, on the average, where $\{B_m\}$ is the set of known members of any other Class B. Similarly, if P is any member of B, then $S|P,\{B_m\}| < S|P,\{A_m\}|$. The two desirable requirements are conveniently combined in the statement of the following problem.

Find the metric, or transformation, of a given class of transformations that maximizes $S|P,\{B_m\}| - S|P,\{A_m\}|$, on the average, if P belongs to Category A, while requiring that the average of $S|P,\{A_m\}| - S|P,\{B_m\}|$ for any P contained in Category B be a positive constant. The constraint of this problem assures that not only points of Category A but also those of B are classified correctly, on the average. The symmetrical situation where $S|P,\{A_m\}| - S|P,\{B_m\}|$ for $P \in B$ is also maximized leads to the same solution.

It is important to note that the above problem is not aimed at maximizing the number of correct decisions. Instead it makes the correct decisions most unequivocal, on the average. It is substantially more difficult to maximize the number of correct classifications. For that purpose a binary function would have to be defined which assumes the more positive of its two values whenever a decision is correct and, conversely, assumes the lower value for incorrect classifications. The sum of this binary function evaluated for each labeled point would have to be maximized. This problem does not lend itself to ready analytical solution; it may be handled, however, by computer methods.

4.2 A SEPARATING TRANSFORMATION

The particular linear transformation that maximizes the mean-square interest distance while holding the sum of the mean-square interset and intraset distances constant is developed below. Recall that the purpose of this transformation is to separate events of dissimilar categories while clustering those that belong to the same class.

The mean-square distance between the M_1 members of the set $\{F_m\}$ and the M_2 members of the set $\{G_p\}$, after their linear transformation, is given in Eq. 53, where f_{ms} and g_{ps} are, respectively, the s^{th} components of the m^{th} and p^{th} members of the sets $\{F_m\}$ and $\{G_p\}$. For the sake of notational simplicity this mean-square interset distance is denoted by $S|\{F_m\},\{G_p\}|$ and is the quantity to be maximized by a suitable choice of the linear transformation. The choice of the notation above is intended to signify that the transformation to be found is a function of the two sets.

$$S(\{F_m\},\{G_p\}) = \frac{1}{M_1M_2} \sum_{m=1}^{M_1} \sum_{p=1}^{M_2} \sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns}(f_{ms} - g_{ps}) \right]^2 \tag{53}$$

The constraint that the mean-square distance θ between points, regardless of the set to which they belong, is a constant, is expressed by Eq. 54, where γ is the coefficient of any point belonging to the union of the sets $\{F\}$ and $\{G\}$, $M_T = \binom{M_1 + M_2}{2}$, and $M = M_1 + M_2$.

$$\theta = \frac{1}{M_T} \sum_{m=1}^{M} \sum_{p=1}^{M} \sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns}(\gamma_{ms} - \gamma_{ps}) \right]^2 = \text{constant } K \tag{54}$$

Both of the above equations may be simplified by expanding the squares as double sums and interchanging the order of summations. Carrying out the indicated operations, we obtain Eqs. 55 and 56.

$$S(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns}w_{nr}x_{sr} \tag{55a}$$

where

$$x_{sr} = x_{rs} = \frac{1}{M_1M_2} \sum_{m=1}^{M_1} \sum_{p=1}^{M_2} (f_{ms} - g_{ps})(f_{mr} - g_{pr}) \tag{55b}$$

and

$$\theta = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns}w_{nr}t_{sr} = K \tag{56a}$$

where

$$t_{sr} = t_{rs} = \frac{1}{M_T} \sum_{m=1}^{M} \sum_{p=1}^{M} (\gamma_{ms} - \gamma_{ps})(\gamma_{mr} - \gamma_{pr}) \tag{56b}$$

The coefficient x_{sr} is the general element of the matrix [X] that is of the form of a covariance matrix and arises from considerations of cross-set distances. The matrix [T] with general coefficient t_{sr}, on the other hand, arises from considerations involving distances between the total number of points of all sets.

We now maximize Eq. 55, subject to the constraint of Eq. 56a, by the method of Lagrange multipliers. Since dw_{ns} is arbitrary in Eq. 57, Eq. 58 must be satisfied.

$$dS - \lambda d\theta = \sum_{n=1}^{N} \sum_{s=1}^{N} dw_{ns} \left[\sum_{r=1}^{N} w_{nr}(x_{sr} - \lambda t_{sr}) \right] = 0 \tag{57}$$

$$\therefore \sum_{r=1}^{N} w_{nr}(x_{sr} - \lambda t_{sr}) = 0 \qquad n = 1, 2, \ldots, N; \; s = 1, 2, \ldots, N \tag{58}$$

Equation 58 can be written in matrix notation to exhibit the solution in an illuminating way. If we let W_n be a vector with N components, $w_{n1} \ldots w_{nN}$, then Eq. 58 may be written as

$$\begin{aligned} &W_1[X-\lambda T] = 0 \\ &\cdots\cdots\cdots \\ &W_n[X-\lambda T] = 0 \\ &\cdots\cdots\cdots \\ &W_N[X-\lambda T] = 0 \end{aligned} \tag{59a}$$

By postmultiplying both sides of the equation by T^{-1}, we obtain Eq. 59b, which is in the form of an eigenvalue problem.

$$\begin{aligned} &W_1[XT^{-1}-\lambda I] = 0 \\ &W_2[XT^{-1}-\lambda I] = 0 \\ &\cdots\cdots\cdots\cdots \\ &W_N[XT^{-1}-\lambda I] = 0 \end{aligned} \tag{59b}$$

Note, that T^{-1} always exists, since T is positive definite. Equations 59a and 59b can be satisfied in either of two ways. Either W_n, the n^{th} row of the linear transformation described by the matrix [W], is identically zero, or it is an eigenvector of the matrix $[XT^{-1}]$. We must make a substitution in the mean-square interset distance given by Eq. 55a in order to find the solution that maximizes S. To facilitate this substitution, we recognize that through matrix notation, Eqs. 55a and 56a can be written as Eqs. 60 and 61.

$$S(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} W_n[X]\, W_n^T \tag{60}$$

$$\theta = \sum_{n=1}^{N} W_n[T]\, W_n^T = K \tag{61}$$

But from Eq. 59a we see that W_nX may always be replaced by λW_nT. Carrying out this substitution in Eq. 60, we obtain Eq. 62, where the constraint of Eq. 61 is also utilized.

$$S(\{F_m\},\{G_p\}) = \lambda \sum_{n=1}^{N} W_n T\, W_n^T = \lambda K \tag{62}$$

It is now apparent that the largest eigenvalue of $[X-\lambda T] = 0$ yields the rows of the transformation that maximizes the mean-square interset distance, subject to the constraint that the mean-square value of all distances is a constant. The transformation is stated by Eq. 63, where $W_1 = w_{11}, w_{12}, \ldots, w_{1N}$ = the eigenvector corresponding to λ_{max}.

$$[W] = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{11} & w_{12} & \cdots & w_{1N} \\ & & \cdots & \\ w_{11} & w_{12} & \cdots & w_{1N} \end{bmatrix} \tag{63}$$

The transformation of this equation is singular, which expresses the fact that the projection of the points along the line of maximum mean-square interset distance and minimum intraset distance is the only important feature of events that determines their class membership. This is illustrated in Fig. 11, where line aa' is in the direction of the first eigenvector of the matrix $[XT^{-1}]$. A point of unknown classification is grouped in Category B because the mean-square difference between its projection on line aa' and the projection of points belonging to set B, $S|P,\{B\}|$, is less than $S|P,\{A\}|$, the corresponding difference with members of set A.

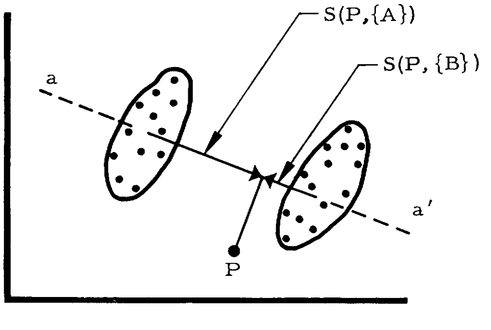


Fig. 11. A singular class-separating transformation.

Forcing the separating transformation to be nonsingular is possible by the imposition of a different constraint on the maximization. Unfortunately, the mathematical difficulty of imposing nonsingularity directly is a formidable task. In general, it requires evaluating a determinant, such as the Gramian, and assuring that it does not vanish. In the following discussion, at first a seemingly meaningless constraint will be imposed on the maximization of the mean-square interset distance. After the solution is obtained, it will be shown that the meaningless constraint can be converted to a constraint that holds the mean-square of all distances constant — the same constraint that we used before.

The mean-square interset distance to be maximized is given by Eq. 55a, which is reproduced here as Eq. 64.

$$S(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns} w_{nr} x_{sr} \tag{64}$$

The constraint that we shall impose is that the mean-square length of the projections of all distances between any pair of points onto the directions W_n be fixed, but in general, different constants. This constraint is expressed by Eq. 65, which differs from the previously used constraint of Eq. 56 only by fixing coordinate by coordinate the mean-square value of all possible distances between points.

$$\sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns} w_{nr} t_{sr} = K_n \qquad n = 1, 2, \ldots, N \tag{65}$$

Assigning an arbitrary constant λ_n to the differential of each of the above N constraints and using the method of Lagrange multipliers in the maximization of S above, we obtain Eq. 66.

$$dS - \sum_{n=1}^{N} \lambda_n dK_n = \sum_{n=1}^{N} \sum_{s=1}^{N} dw_{ns} \left[\sum_{r=1}^{N} w_{nr}(x_{sr} - \lambda_n t_{sr}) \right] = 0 \tag{66}$$

When we make use of the convenient matrix notation employed earlier, we obtain Eq. 67, which differs significantly from Eq. 59a, despite the similar appearance of the two equations.

$$\begin{aligned} &W_1[X-\lambda_1 T] = 0 \\ &W_2[X-\lambda_2 T] = 0 \\ &\cdots\cdots\cdots \\ &W_N[X-\lambda_N T] = 0 \end{aligned} \tag{67}$$

The solution of Eq. 67 states that each row of the linear transformation, W_n, is a different eigenvector of the $[XT^{-1}]$ matrix. The transformation $[W]$ is therefore orthogonal. Equation 68 is a further constraint that converts that of Eq. 67 to holding the mean-square of all distances constant, and thus accomplishes our aim.

$$K = \sum_{n=1}^{N} K_n \tag{68}$$

Note that before we knew that the rows of the transformation [W] would be orthogonal, the condition expressed by Eq. 68 did not fix the total distances. The procedure above resulted in finding the nonsingular orthogonal transformation that optimally separates the classes and optimally clusters members of the same class.

We shall now compute the mean-square interset distance S of Eq. 64. To facilitate the computation, S will be written in matrix notation:

$$S(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} W_n X W_n^T \tag{69}$$

From Eq. 67 it is seen, however, that if S is maximum, $W_n X$ may be replaced with $\lambda_n W_n T$ to obtain Eq. 70 from Eq. 65 (in matrix notation)

$$S_{max}(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} \lambda_n W_n T W_n^T \tag{70}$$

where $W_n T W_n^T = K_n$. Equation 71 is thus obtained. It is now readily seen, with reference to Eq. 62, that the upper bound on the mean-square interset distance is achieved by the singular transformation discussed earlier, and we pay for forcing the transformation to be nonsingular by achieving only a reduced separability of classes.

$$S_{max}(\{F_m\},\{G_p\}) = \sum_{n=1}^{N} \lambda_n K_n \tag{71}$$

Before leaving the discussion of class-separating transformations, a few important facts must be pointed out. A simple formal replacement of the matrices X and T by other suitably chosen matrices yields the solution of many interesting and useful problems. It is not the purpose of the following remarks to catalog the problems solved by the formal solution previously obtained; yet some deserve mention because of their importance. It may be readily verified, for instance, that replacing T by I is equivalent to maximizing the between-set distances, subject to the condition that the volume of the space is a constant. The transformation that accomplishes this is orthogonal with rows equal to different eigenvectors of the matrix X. This is a physically obvious result, of course, since the eigenvectors of X are the set of orthogonal directions along which interset distances are maximized, on the average. A figure that would illustrate the result is very similar to Fig. 1.

Another replacement that must be singled out is the substitution of the matrix L for T, where L is the covariance matrix associated with all intraset distances (distances among like events). Eigenvectors of $[X-\lambda L]$ form rows of the transformation that maximizes interset distances while holding intraset distances constant. This problem is essentially the same as the maximization of interset distances while holding the sum of interset and intraset distances constant, yet the relative separation of sets achieved by the two transformations is different. The difference may be exhibited by computing the ratio of the mean-square separation of sets to the mean clustering of elements within the same set, as measured by the mean-square intraset distance. It can be concluded, therefore, that the constraint employed in the maximization of interset distances does have an influence on the degree of separation achieved between sets.

Throughout this section the class-separating transformations were developed by reference to the existence of only two sets, $\{F_m\}$ and $\{G_p\}$. The results obtained by these methods are more general, however, because they apply directly to the separation of an arbitrary number of sets. For instance, in the maximization of the mean-square interset distance, there is no reason why the matrix X should involve interset distances between only two sets. An arbitrary number of sets may be involved, and the interset distances are simply all those distances measured between two points not in the same set. Similar arguments are valid for all the other matrices involved. The only precaution that must be taken concerns the possible use of additional constraints specifying preferential or nonpreferential treatment of classes. These additional constraints may take the form of requiring, for instance, that the mean-square intraset distance of all sets be equal or be related to each other by some constants. Aside from these minor matters, the results apply to the separation of any number of classes.

4.3 MAXIMIZATION OF CORRECT CLASSIFICATIONS

The correct classification of points of the set F are made more unequivocal by the linear transformation that makes any event F_n of set F more similar to members of F, on the average, than to those of another set G. One of the ways in which the average unequivocalness of correct classificatory decisions may be stated mathematically is to require that a numerical value associated with the quality of a decision be maximized, on the average. Of the several quantitative measures of the quality of a decision that may be defined, one that readily lends itself to mathematical treatment is given in Eq. 72. The difference in the similarity between a point P and each of the two sets, F and G, is a quantity Q, which increases as the decision regarding the classification of P becomes more unequivocal.

$$S(P,\{G_m\}) - S(P,\{F_m\}) = Q \tag{72}$$

Since decisions in previous sections were based on the comparison of Q with a suitable threshold value (such as zero), we now want to find that linear transformation that maximizes Q, on the average, whenever Q is to be positive. If P is a member of the set F, then P is closer to F than to G, and thus Q is to be positive. The maximization of Q for P ϵ F results in maximizing the margin with which correct decisions are made, on the average. The foregoing maximization is stated in Eq. 73, subject to the constraint expressed by Eq. 74. The latter simply states that if P ϵ G, the average decision is still correct, as measured by the margin K.

$$\overline{S(F_n,\{G_p\}) - S(F_n,\{F_m\})}^{\,n} = \overline{Q} = \text{maximum} \tag{73}$$

subject to

$$\overline{S(G_n,\{F_m\}) - S(G_n,\{G_p\})}^{\,n} = \overline{K} = \text{constant} > 0 \tag{74}$$

Maximization of $\overline{Q} + \overline{K}$ has the same solution.

By utilizing previously obtained results, these equations are readily solved for the optimum linear transformation. Rewriting the first term of Eq. 73, we note that it expresses the mean-square interset distance between sets F and G and may be written as in Eq. 75, where Eq. 53 and the simplifying notation of Eq. 55 are employed.

$$\overline{S(F_n,\{G_p\})}^{\,n} = S(\{F_n\},\{G_p\}) = \frac{1}{M_1M_2}\sum_{m=1}^{M_1}\sum_{p=1}^{M_2}\sum_{n=1}^{N}\left[\sum_{s=1}^{N} w_{ns}(f_{ms}-g_{ps})\right]^2 \qquad (75a)$$

$$\overline{S(F_n,\{G_p\})}^{\,n} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}w_{nr}x_{sr} \qquad (75b)$$

The second term of Eq. 73 is the mean-square intraset distance of set F and can be expressed as in Eq. 76. The argument of the covariance coefficient $u_{sr}(F)$ signifies that it is a covariance of elements of the set F.

$$\overline{S(F_n,\{F_m\})}^{\,n} = S(\{F_n\},\{F_m\}) = \frac{1}{(M_1-1)\,M_1}\sum_{p=1}^{M_1}\sum_{m=1}^{M_1}\sum_{n=1}^{N}\left[\sum_{s=1}^{N} w_{ns}(f_{ps}-f_{ms})\right]^2 \qquad (76a)$$

$$\overline{S(F_n,\{F_m\})}^{\,n} = \frac{2M_1}{M_1-1}\sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}w_{nr}\,u_{sr}(F) \qquad (76b)$$

Similarly, the first term of Eq. 74 is the mean-square interset distance, and the second term is the intraset distance of set G. The maximization problem can thus be restated by Eqs. 77a and 77b.

$$\text{maximize } \overline{Q} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}w_{nr}\left[x_{sr} - \frac{2M_1}{M_1-1}\,u_{sr}(F)\right] \qquad (77a)$$

subject to

$$\overline{K} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}w_{nr}\left[x_{sr} - \frac{2M_2}{M_2-1}\,u_{sr}(G)\right] \qquad (77b)$$

Following the methods used earlier, we can write the solution of this problem by inspection.

$$d\overline{Q} - \lambda d\overline{K} = \sum_{n=1}^{N}\sum_{s=1}^{N} dw_{ns}\left[\sum_{r=1}^{N} w_{nr}\left\{x_{sr} - \frac{2M_1}{M_1-1}\,u_{sr}(F) - \lambda\left(x_{sr} - \frac{2M_2}{M_2-1}\,u_{sr}(G)\right)\right\}\right] \qquad (78)$$

From Eq. 78 it follows that Eq. 79a must hold, where α_{sr} and β_{sr} are given by Eqs. 79b and 79c.

$$\sum_{r=1}^{N} w_{nr}(\alpha_{sr} - \lambda\beta_{sr}) = 0 \qquad n = 1, 2, \dots, N; \; s = 1, 2, \dots, N \tag{79a}$$

$$\alpha_{sr} = x_{sr} - \frac{2M_1}{M_1 - 1} u_{sr}(F) \tag{79b}$$

$$\beta_{sr} = x_{sr} - \frac{2M_2}{M_2 - 1} u_{sr}(G) \tag{79c}$$

By reference to earlier results, such as those expressed by Eq. 58, the transformation whose coefficients w_{ns} satisfy an equation of the presiding form is the solution of the eigenvalue problem of Eq. 80, where W_n is a row of the matrix expressing the linear transformation.

$$\begin{aligned} W_1[\alpha-\lambda\beta] &= 0 \\ W_2[\alpha-\lambda\beta] &= 0 \\ \cdots&\cdots \\ W_N[\alpha-\lambda\beta] &= 0 \end{aligned} \tag{80}$$

Analogous to the arguments used in the previous section, the above solution yields a singular transformation. Forcing the transformation to be nonsingular, in the manner already outlined, results in the optimum transformation as an orthogonal transformation, where each row of the matrix $[W]$ is an eigenvector of $[\alpha-\lambda\beta] = 0$. Furthermore, it is readily shown that the solution so obtained indeed maximizes $\overline{Q}$.

It is interesting to note that the maximization of the average correct classifications can be considered as the maximization of the difference between interset and intraset distances. This alternate statement of the problem can be exhibited by the addition of Eq. 77b to Eq. 77a.

$$\overline{Q} + \overline{K} = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns} w_{nr} \left[2x_{sr} - \left\{ \frac{2M_1}{M_1 - 1} u_{sr}(F) + \frac{2M_2}{M_2 - 1} u_{sr}(G) \right\} \right] \tag{81}$$

But the expression within the braces is simply the covariance ℓ_{sr} associated with all intraset distances. Since K is a constant, the maximization of Eq. 81 is equivalent to the maximization of $\overline{Q}$.

In summing up the results of this section, we see that the problem of learning to measure similarity to events of a common category, while profiting from knowledge of nonmembers of the same category, can be treated as a maximization or minimization problem. A metric, or a linear transformation, is found from a class of metrics, or transformations, that solves mathematical problems that express the desire not only to cluster events known to belong to the same category but also to separate those that belong to different categories. Within the restricted class of metrics, or transformations, considered in this section, the solutions are in the form of eigenvalue problems which emphasize features that examples of a category have in common, and which at the same time differ from features of other categories.

V. NONLINEAR METHODS IN CLASSIFICATORY ANALYSIS

A number of different ideas that deal with nonlinear methods in classificatory analysis will now be discussed. First, we consider the enlargement of the class of transformations so that the distribution of similar events may be altered in a larger variety of ways. Next, we present another geometric interpretation of classification that gives rise to a slightly different decision rule from that used before. The relationship between the two is explored. Finally, we present a set of miscellaneous unsolved problems and thoughts that represent directions that may be taken by continuing research.

5.1 ENLARGING THE CLASS OF TRANSFORMATIONS

The central concept motivating the development of this theory is that there is a transformation that exhibits the category defining common attributes of members of a set, and that this transformation can be found from the solution of a problem that optimizes, with respect to the choice of a specific transformation, some property of a set of known events. In previous sections, several different properties of the set of known events were investigated in regard to their application as optimization criteria. Clustering of points of the same set, maximal separation of events of different sets, and maximization of the unequivocalness of a classificatory decision are only a few of the criteria to note. The principal limitation of the results obtained from these considerations stems not from the optimization criteria but from the limitation on the class of transformations among which the optimum was sought.

Achieving success with the linear methods explored thus far requires that the distribution of like events be well fitted by an ellipsoidal boundary, or, at least, ellipsoidal boundaries fitted to different categories should not intersect. The methods developed in previous sections describe the linear transformation, or the error criterion with which distance is measured, which maximizes the correlation between like events and minimizes it between unlike ones. It is clear, however, that for many categories high correlation may not exist between like events even after their linear transformation. Successful classification even of categories of this type may be achieved, if the class of transformations among which the optimum is sought is enlarged. We recall that the class of metrics given in Eq. 82 could be viewed as the application of the mean-square error criterion after a linear transformation of the events. With use of the notation often employed in engineering, the metric of Eq. 82 can be written as in Eq. 83 and can be instrumented by the system of Fig. 12.

$$d^2(a,b) = \sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns}(a_s - b_s) \right]^2 \tag{82}$$

In Eq. 83, $W(t, \tau)$ is a linear function of two variables, and the methods previously described simply find the best one that minimizes the output, on the average, or performs some similar but preselected desirable operation.

$$d^2(a, b) = \int \left[\int W(t, \tau)[a(t)-b(t)]\, dt \right]^2 d\tau \tag{83}$$

There are many categories of inputs whose members cannot be made "close" by any linear network. Suppose, for instance, that a and b belong to the category of waveforms that have 5 zero crossings. Since the member waveforms of this category are close to each other in the sense of having the same number of zero crossings and since counting zero crossings is not a linear operation, it is evident that the methods described will fail to learn the correct recognition of membership in the category of interest. On the other hand, it is quite possible to employ linear methods to solve the problem successfully if inputs are represented differently; that is, if the model of the physical world from which the waveforms were obtained is altered. In this example the waveform acts as its own model. In an arithmetized form, on the other hand, the waveform can be represented by the coefficients of its expansion in one of several suitable generalized harmonic functions. It can also be represented by the results of certain tests that can be performed on the waveform. Linear methods can be forced to solve problems of the type described, if the selection of the model of the physical world includes tests on which a linear function can be found which are such that, as measured by the linear functional, tests on different members of a category have similar outcomes. An extreme example of such a test is the examination of the waveform in order to establish the number of times its magnitude is less than ϵ. In the example above, waveforms with 5 zero crossings would yield very similar results when subjected to this test, and no linear functional at all is necessary to exhibit this property. The choice of the model, the nature of the harmonic expansion, or the nature of the tests is an ad hoc choice on the part of the designer. In this sense, therefore, the designer determines the class of basically nonlinear problems that can be solved with linear methods. In the word recognition example of Appendix C, for instance, it was an ad hoc choice based on past experience to represent speech by sonograph recordings. Many of the other possible representations of speech, such as the successive samples of the microphone output voltage, could have resulted in the failure of linear methods to learn the categories.

The degree of success achieved by linear methods depends on the choice of the model and the nature of the sets to be categorized. This dependence may be eliminated by

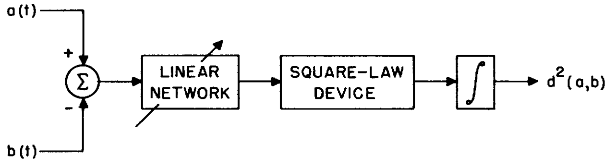


Fig. 12. Instrumentation of a class of metrics.

enlarging the class of transformations within which we search for the optimum one. In the categorization of waveforms of 5 zero crossings, for instance, the correct category definition may be obtained if we do not restrict the network of Fig. 12 to be a linear network. If the network contains nonlinear elements, such as flip-flops, and the like, it is at least plausible that "having equal numbers of zero crossings" is a category that can be learned.

In the following we would like to consider a class of nonlinear transformations. Whereas before we considered the class obtained by measuring the Euclidean distance on a linear transformation of a set of points, we now consider Euclidean distance measured on a continuous nonlinear transformation of the points. This type of transformation describes the operation of stretching and compressing a rubber sheet in an arbitrary manner to bring members of a set closest to each other.

All of the minimization problems of the preceding sections can be extended and carried out for nonlinear transformations as well. In the following section only two of these will be derived as illustrative examples. The first minimization problem is an extension of the method of clustering members of a category after their nonlinear transformation. The second problem is an extension of the method of clustering members of sets while separating sets. This problem is solved — within the class of linear transformations — in Section IV.

The class of continuous transformations is too general to yield a practical solution. In the discussion that follows, therefore, the class of transformations within which a solution is sought will be restricted to a certain polynomial transformation.

Problem 1. A Clustering Transformation

Let us assume that each coordinate of the N-dimensional space undergoes a continuous transformation given by the polynomial of Eq. 84, which maps the origin into itself.

$$y_n = \sum_{p=1}^{K} a_{np} x_n^p \tag{84}$$

This transformation is illustrated in Fig. 13, which shows that a set of equally spaced grid lines under transformation can be spaced in an arbitrary manner along each coordinate. In this way the N-dimensional space can be locally compressed or expanded. The square of the Euclidean distance between two points F'_m and F'_s obtained from the transformation of F_m and F_s, respectively, is expressed by Eq. 85.

$$d_e^2(F'_m, F'_s) = \sum_{n=1}^{N} \left[\sum_{p=1}^{K} a_{np} \left(f_{mn}^p - f_{sn}^p \right) \right]^2 = d^2(F_m, F_s) \tag{85}$$

It is readily proved that Eq. 85 is indeed a metric if the transformation expressed by Eq. 84 is single-valued and one-to-one. If we allow the folding of the space by not

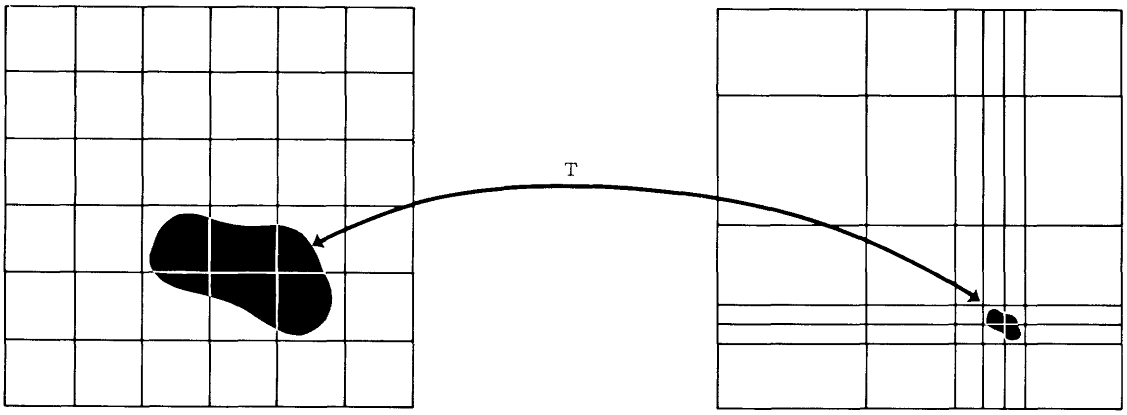


Fig. 13. A nonlinear transformation.

restricting the coefficients of the polynomial, we have a multivalued transformation; and we no longer have a metric because $d(F_m, F_s)$ may be zero even in $F_m \neq F_s$. The Euclidean distance, of course, can be made smaller between two points without the restriction, and therefore $d_e^2(F'_m, F'_s)$ is a useful measure of distance on the transformed space.

The problem is to find the particular nonlinear transformation of the type discussed above which minimizes the mean-square distance between a set of points of a given category after their transformation. In order to obtain a unique solution, we impose the quite arbitrary constraint that a specific point $x_o = (x_{o1}, x_{o2}, \ldots, x_{on})$ should be mapped into another specified point, $y_o = (y_{o1}, y_{o2}, \ldots, y_{on})$.

The mean-square distance after transformation is given by Eq. 86, where M is the number of elements in the set.

$$\overline{d_e^2(F'_m, F'_s)}^{s,m} = \frac{2M}{(M-1)} \sum_{n=1}^{N} \sum_{p=1}^{K} \sum_{r=1}^{K} a_{np} a_{nr} \left(\overline{f_n^p f_n^r} - \overline{f_n^p}\,\overline{f_n^r}\right) \tag{86}$$

The notation can be simplified by letting Eq. 87 express the simplifying substitution that yields Eq. 88a. In matrix notation this can be expressed by Eq. 88b.

$$\overline{f_n^p f_n^r} - \overline{f_n^p}\,\overline{f_n^r} = u_{pr}(n) = u_{rp}(n) \tag{87}$$

$$\overline{d_e^2(F'_m, F'_s)}^{s,m} = S(\{F'_m\}, \{F'_s\}) = \frac{2M}{M-1} \sum_{n=1}^{N} \sum_{p=1}^{K} \sum_{r=1}^{K} a_{np} a_{nr} u_{rp}(n) \tag{88a}$$

$$S(\{F'_m\}, \{F'_s\}) = \frac{2M}{M-1} \sum_{n=1}^{N} a_n [U(n)]\, a_n^T \tag{88b}$$

The constraint that x_o map into y_o is expressed by Eq. 89, which contains N different constraints, one for each component of the mapping.

$$y_{on} = \sum_{p=1}^{K} a_{np} x_{on}^{p} \qquad n = 1, 2, \ldots, N \tag{89}$$

By defining the vector z_n in Eq. 90, the above constraint may be readily expressed in vector notation.

$$z_n = \left(x_{on}, x_{on}^{2}, \ldots, x_{on}^{p}, \ldots, x_{on}^{k} \right) \tag{90}$$

$$y_{on} = a_n I z_n^T = a_n \cdot z_n \qquad n = 1, 2, \ldots, N \tag{91}$$

By using the method of Lagrange multipliers, S can be minimized subject to the constraints of Eq. 91:

$$ds - \sum_{n=1}^{N} \lambda_n dy_{on} = 0 \tag{92}$$

In a manner similar to the methods employed earlier, the solution of the minimization problem is expressed by Eq. 93, which can be solved for the vector a_n.

$$a_n[U(n)] - \lambda_n z_n = 0 \qquad n = 1, 2, \ldots, N \tag{93a}$$

$$a_n = \lambda_n z_n [U^{-1}(n)] \qquad n = 1, 2, \ldots, N \tag{93b}$$

The constant of λ_n can be evaluated by substituting Eq. 93b into Eq. 91 and solving for λ_n.

$$y_n = a_n z_n^T = \lambda_n z_n [U^{-1}(n)]\, z_n^T \tag{94a}$$

$$\therefore \lambda_n = \frac{y_n}{z_n [U^{-1}(n)]\, z_n^T} \tag{94b}$$

Substituting the value of λ_n into Eq. 93b, we obtain the coefficients of the transformation.

$$a_n = \left(\frac{y_n}{z_n [U^{-1}(n)]\, z_n^T} \right) z_n [U^{-1}(n)] \tag{95}$$

The mean-square distance S can now be computed by substituting Eq. 95 in Eq. 88b:

$$S_{min} = \frac{2M}{M-1} \sum_{n=1}^{N} \left(\frac{y_n}{z_n [U^{-1}]\, z_n^T} \right)^2 z_n [I] \left(z_n [U^{-1}(n)] \right)^T \tag{96a}$$

which may be simplified as

$$S_{min} = \frac{2M}{M-1} \sum_{n=1}^{N} \left(\frac{y_n}{z_n [U^{-1}(n)]\, z_n^T} \right)^2 z_n [U^{-1}(n)]^T\, z_n^T \tag{96b}$$

Note that $[U^{-1}(n)]^T = [U^{-1}(n)]$, since $[U^{-1}(n)]$ is a symmetric matrix. This fact allows further simplification of S_{min} and it can be shown that the extremum thus found is indeed a minimum.

$$S_{min} = \frac{2M}{M-1} \sum_{n=1}^{N} \frac{y_n^2}{\left(z_n[U^{-1}(n)]z_n^T\right)} \tag{97}$$

This transformation may be used to further cluster a set of points that have been maximally clustered by linear transformations.

Problem 2. A Class-Separating Transformation

Let us assume that each coordinate of the N-dimensional space undergoes a continuous transformation given by the polynomial of Eq. 84. The object of this problem is to find that particular nonlinear transformation that minimizes the mean-square distance between points in the same set, while keeping the distance between the means of the sets a constant. Let us assume here, as in Section IV, that there are only two categories, F and G. We see from Eqs. 88a and 88b that the mean-square distance between points in the same set is the quantity Q expressed in Eq. 98.

$$\begin{aligned} Q &= \overline{d_e^2(F'_m, F'_s)}^{s,m} + \overline{d_e^2(G'_m, G'_s)}^{s,m} \\ &= \frac{2M_1}{M_1-1} \sum_{n=1}^{N} a_n[U_F(n)]\, a_n^T + \frac{2M_2}{M_2-1} \sum_{n=1}^{N} a_n[U_G(n)]\, a_n^T \end{aligned} \tag{98}$$

where M_1 and M_2 are the number of given samples of F and G, respectively. The other symbols have the same meanings as in the preceding problem. The simplification of Eq. 98 results from the definition of the matrix $\hat{U}(n)$, given in Eq. 99b.

$$Q = \sum_{n=1}^{N} a_n[\hat{U}(n)]\, a_n^T \tag{99a}$$

where

$$[\hat{U}(n)] = \frac{2M_1}{M_1-1}[U_F(n)] + \frac{2M_2}{M_2-1}[U_G(n)] \tag{99b}$$

The distance between the means of the transformed sets of points is given in Eq. 100 and is denoted by the constant K.

$$\begin{aligned} K &= \sum_{n=1}^{N} \left[\sum_{p=1}^{K} a_{np}\left(\overline{f_n^p} - \overline{g_n^p}\right)\right]^2 \\ &= \sum_{n=1}^{N} \sum_{p=1}^{K} \sum_{q=1}^{K} a_{np}a_{nq}\left(\overline{f_n^p} - \overline{g_n^p}\right)\left(\overline{f_n^q} - \overline{g_n^q}\right) = \sum_{n=1}^{N} a_n[B(n)]\, a_n^T \end{aligned} \tag{100}$$

where $[B(n)]$ is a matrix that has a general element b_{pq} given by

$$b_{pq} = b_{qp} = \left(\overline{f_n^p} - \overline{g_n^p}\right)\left(\overline{f_n^q} - \overline{g_n^q}\right) \tag{101}$$

The minimization of Q subject to the constraint K can be carried out by the method of Lagrange multipliers, and results in the familiar solution given in Eq. 102.

$$\min(Q-\lambda K) \rightarrow a_n[\hat{U}(n)-\lambda B(n)] = 0 \tag{102}$$

The optimum nonlinear transformation that clusters members of the sets while it keeps the sets separated is a polynomial transformation in which the polynomial coefficients are components of the eigenvector corresponding to the smallest eigenvalue of Eq. 102.

5.2 GEOMETRICAL INTERPRETATION OF CLASSIFICATION

The geometrical interpretation of making decisions of classification is that the vector space in which events are represented by points must be divided into nonintersecting regions, one region corresponding to each of the categories. A decision of classification consists of assigning to the event to be classified the name of the category associated with the region in which the event is located.

The decision procedure of the preceding sections constructed the boundaries of these regions and the notion of "inside" versus "outside". The optimum size, shape, and location for the boundaries was derived within the possible shapes that the class of transformations considered at the time was able to generate.

The objective of the following discussion is to present a different geometrical interpretation of making classification decisions and to show the relationship between the two interpretations.

In order to recognize membership of a vector in a class represented by a set of sample vectors, we want to find some invariant property of the set of samples. Let the invariant property be $u_1(v_1, v_2, \ldots, v_N)$, some function of the vector v. Invariance is understood to mean that the function $u_1(v)$ will have substantially the same value, say K_1, whenever v is a member of class F. Similarly, $u_2(v)$ is another function of the vector v which is such that whenever v is a member of another class, G, $u_2(v)$ will have substantially the same value, say K_2, but $K_2 \neq K_1$. Since, according to our original assumptions, any arbitrary point v can be the member of only one class, we can assume a function u(v) which is such that whenever v is contained in class F, $u(v) \approx K_1$ and whenever v is contained in class G, $u(v) \approx K_2$. The function u(v) is a surface over the multidimensional space, and it is so constructed that along the u axis known samples of the different classes fall into disjointed, nonoverlapping intervals. Figure 14 illustrates this situation for a two-dimensional vector space, where $u(v_1, v_2)$ is a three-dimensional surface. The heights of the surface over samples of class F are highly clustered, and the classes along $u(v_1, v_2)$ are separable. It is readily appreciated that

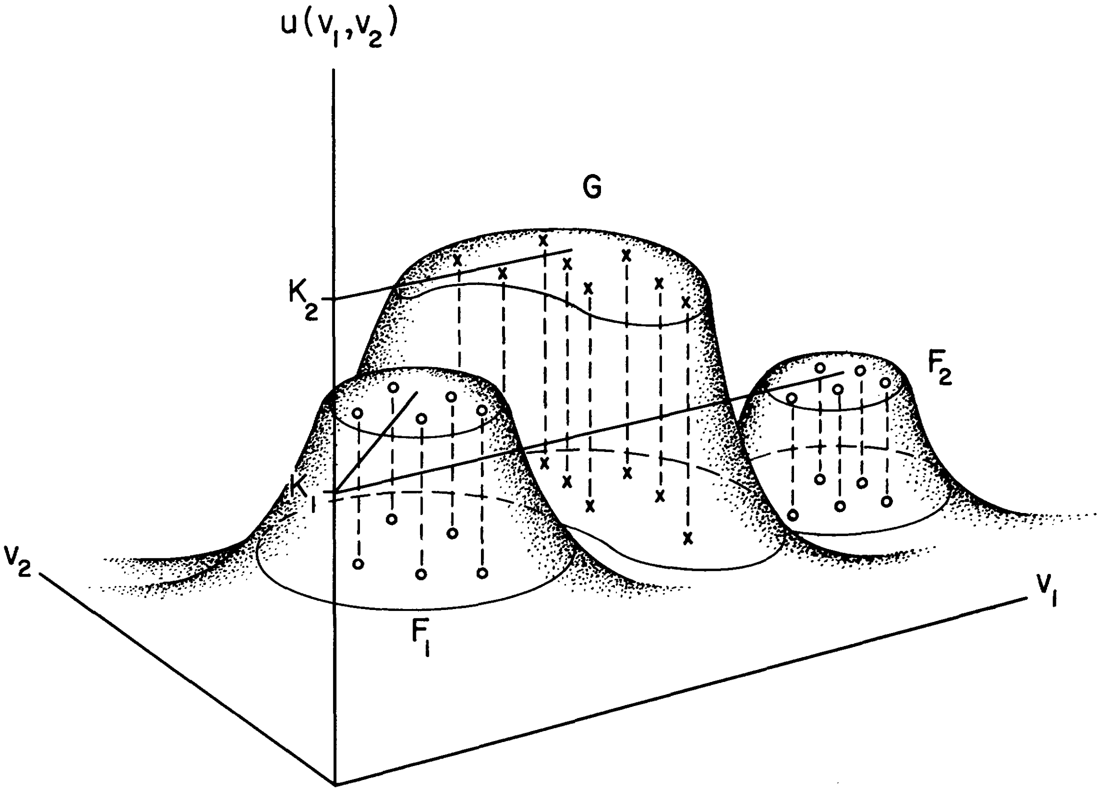


Fig. 14. Under transformation $u(v_1, v_2)$, F and G are separable.

regardless of the manner in which points of the sets F and G are distributed, a surface can always be constructed that clusters members of F and members of G along the height of the surface and keeps the two clusters apart. Furthermore, it does not matter that a class has several disjointed subclasses, as shown in Fig. 14. In spoken word recognition, for example, F_1 may represent male utterances and F_2 may represent female utterances of the same word; yet the function u(v) has the same value over both F_1 and F_2.

Now, we shall show that minimization of the mean-square distance between members of the same class after a transformation of the vector space is equivalent to the process of constructing a surface u(v) in such a way that it has a minimum variance about the mean over members of the same class.

The mean-square distance after a transformation u(v) is given in Eq. 103a. Squaring and averaging with respect to m and n yield Eq. 103b, where $\sigma_u^2(F)$ stands for the variance after transformation u(v) of the set of points contained in F.

$$\overline{d^2(F'_m, F'_n)}^{m,n} = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}[u(F_m)-u(F_n)]^2 \qquad (103a)$$

$$= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\left[u^2(F_m)-2u(F_m)u(F_n)+u^2(F_n)\right] = 2\,\overline{u^2(F)} - 2\,\overline{u(F)}^2 = 2\sigma_u^2(F)$$

(103b)

The class of functions u within which the minimization is carried out again limits the nature of the distributions of F and G samples that can be successfully classified. In particular, if the function u(v) is of the form given in Eq. 104, then, by suitable choice of the degree K, a completely general nonlinear function can be arbitrarily closely approximated in a finite region of the vector space.

$$u(v_1, v_2, \ldots, v_N) = \sum_{n=0}^{K} \cdots \sum_{j=0}^{K} \sum_{i=0}^{K} a_{ij\ldots n} v_1^i v_2^j \ldots v_N^n \qquad (104)$$

The decision rule of the previous sections can now be given another simple geometrical interpretation. A decision consists of evaluating the height of the surface over the vector to be classified, u(v), and comparing this height with the average heights over the two sets of given samples. The decision is that v is a member of that class to which it is closer, on the average, after transformation u.

It is shown in Appendix B that the decision rule of the preceding sections is equivalent to Bayes' rule, if the transformation is linear and the probability densities of the sets of N-dimensional vectors are Gaussian. It has been shown by my colleague R. Hines, that if the transformation is allowed to be completely arbitrary (but piecewise continuous) then this decision rule is equivalent to Bayes' rule even for arbitrary probability densities of the classes F and G. The proof was based on certain arguments of mathematical logic. A shorter proof employing calculus of variations is given in Appendix D.

5.3 SOME UNSOLVED PROBLEMS

The preceding mathematical developments were based on the assumption that sets of noise-free examples of the categories are given and that membership determining functionals must be developed from these examples. There are many practical instances, however, when noise-free samples of the category are not available. In these instances the observed differences between samples are not solely caused by genuine, permissible variants of the class, but are affected by the corrupting influence of noise. In the classification of waveforms of 5 zero crossings, for example, the presence of noise may cause a change in the number of zeros, a fact that has serious consequences in learning the correct category definition. In another example of practical interest, this effect is even more pronounced. Consider the automatic recognition of the acoustic signals caused by a given type of motor vehicle and its differentiation from those of other types. The situation may arise in which the vehicle types of interest may be observed only in a noisy environment in which many types of systematic and random noises corrupt the desired signals. These noises affect the decision regions determined by the learning process and thus result in an incorrect definition of the category.

Quite a different situation exists if learning is done on uncorrupted samples of the category, but events to be classified exist in a noisy environment. In this case the problem becomes that of detecting a class of signals in noise. Neither the problem of learning on noisy examples nor that of recognizing in the presence of noise has been solved, although the latter is expected to have a fairly straightforward solution.

Another unsolved problem not yet sufficiently investigated concerns the convergence of the "learning" process to the identification of the correct category. It must be realized, of course, that no statement can be made about convergence to the correct category unless something can be said about how examples on which we learn are chosen from the set of all possible members of the various categories. The question then remains of whether it is possible to specify the method of selection of examples sufficiently to prove convergence and yet insufficiently enough to leave the category undetermined. It is doubtful that this is possible. Consider the following hypothetical example as motivation of this doubt. Suppose that it can be shown that convergence to the correct category necessitates that the examples be uniformly distributed over the region of signal space occupied by all members of the category. It then seems that the examples, together with the knowledge of their distribution over the set of all members of the category, serve as a definition of the category, and make it unnecessary to try to "learn" it. Although of doubtful value, convergence must be studied, for the reward is great if a solution is reached. In most practical instances in which pattern recognition is employed there is no way of assuring adherence to the method of selecting examples, even if they could statistically or otherwise be specified. Convergence, therefore, should be studied perhaps only empirically and be considered primarily in the comparison of categorization schemes.

The methods of categorization described in the preceding sections are deterministic. Given a set of examples of a category, the machine's performance can be determined exactly. In the remaining portion of this section the utilization of already learned concepts or categories in a more efficient classification of new events will be described. This application results in machine performance on a set of inputs that is not predictable solely from knowledge of the machine's circuit diagram and examples of the category in question. As a result of exposure to a set of examples, the machine forms a concept of the category through the construction of the optimum metric that measures membership in the category. This metric expresses, and is capable of measuring, similarity to an already learned category. Suppose, for example, that the category "square" has been learned by the presentation of a set of examples of squares; and, as a result, similarity to a square, or squareness, can be expressed quantitatively. One can now consider the introduction of the metric measuring squareness as a new dimension of the signal space. Squareness is a new measurable property of the environment; it is a test that has a quantitative outcome of the same general type as the tests on the environment that are performed by the original N dimensions of the signal space. The interesting consequence of introducing an already learned concept (squareness) as a new dimension of the signal space is that the vocabulary of the machine is thereby enlarged. The machine is able to describe its environment in a language that depends on the machine's previous experience. Presented with the same set of examples on which to learn a given category, two identical machines with different prior experiences will react in different ways. In some respects this is a phenomenon not unlike that which we observe in human behavior.

VI. GENERAL REMARKS ON PATTERN RECOGNITION

We would now like to present some philosophical ideas about the nature of pattern recognition and to draw some conclusions regarding the relationship between pattern recognition as a whole and this work in particular. No attempt will be made to describe or even to comment briefly on individual works. Such an attempt is not within the scope of this report. Guide lines for future work in pattern recognition and objectives to be achieved, however, are discussed, since these seem to follow from the methods developed earlier in our report.

One of the chief difficulties confronting the reader of literature in pattern recognition and in artificial intelligence is the absence of a common terminology. This situation is quite understandable if we consider the wide range of problems tackled by researchers of varied backgrounds. Engineers, mathematicians, psychologists, and neurophysiologists work on problems dealing with character recognition, speech or speech sound identification, automatic chess programming, mathematical theorem proving, and machine translation of languages, to mention only a few of the research areas.

Although the subject of this report is classification decisions in pattern recognition, the word "pattern" has not been defined. No definition was given, partly because none was needed and partly because a better understanding of the term may be gained now, after the subject has been developed. A pattern can be defined as the set of invariants of a class of signals. The invariants are those common features of the samples of the class which have substantially the same numerical values regardless which sample is singled out for examination. Thus we can define a pattern as those features of a set of things in which members of the set are close.

There are two types of problems to be distinguished in pattern recognition. One is the detection and extraction of the common pattern of a set of things through a process of learning; the other is the recognition of the presence of an already learned pattern in the set of things to be classified. Most of the special devices built to recognize different printed or handwritten characters, speech sounds, or things belonging to other categories accomplish the task by recognizing the presence or absence of an already known set of parameters in the input. In this sense these special devices operate in a manner not unlike a switching circuit that makes decisions based completely on a "truth table" built in by design. Learning in these devices is accomplished by the designer and not by the machine; in fact, the machine's every operation in recognizing membership in a category is spelled out in advance. It is for this reason that, no matter how ingenious these machines may be, their design and development should not be confused with fundamental research in pattern recognition, which should concern itself with machine learning methods for automatically detecting the pattern. For a long time to come, machines that recognize patterns learned and defined by human beings will demonstrate superior performance over those that use machine learning to establish a definition of membership in a category, since intelligent human beings learn better than the best of the machines of the near future. It is equally true that those methods of recognition that use patterns fixed by design cannot solve general problems but only those for which the fixed pattern expresses correctly the invariants of the class of similar signals. Machines that learn are at least potentially capable of handling problems of far greater variety.

It is apparent, upon inspecting the fundamental differences between methods that recognize a known pattern and those that first learn it, that in the former the error criterion with which similarity is judged is fixed by design, but in the latter only the rules for arriving at an error criterion are determined. Although the second kind of machine may very well be deterministic inasmuch as the output may be specified if the input is known, this fact does not preclude its possessing a certain capability for learning. In the method of categorization discussed in the preceding sections, prior to giving the set of inputs, we do not specify what constitutes membership in a category or how to measure membership in it. The method does specify, however, how to obtain the metric from the input with which membership in the category is measured. This method of operation is quite different from correlation, for example, where it is determined that similarity is always measured by closeness in the mean-square sense. The inflexibility of fixing the error criterion in this way is reasonable only if there are other grounds on which a human may base such a decision. Learning, however, in that case, is done by the human and not by the machine.

It is important that the error criterion be determined by the set of labeled examples of the categories that are to be learned. If this were not done, the error criterion would have to be built in by design and permanently fixed. This would lead to the unsatisfactory situation of fixing the "sense" in which the things in a set are similar without first inspecting them.

On the basis of the foregoing discussion, it is suggested that a meaningful way of subdividing the field of pattern recognition should be based on whether the system under consideration has a built-in error criterion for measuring similarity or whether such a measure is developed through a built-in rule by the input itself. The systems that develop error criteria can, in some respect, be considered a self-organizing or self-adapting systems.

Besides the error criterion with which similarity is judged, the parallel or the sequential processing of input information often receives a great deal of attention. There are valid arguments in favor of both. The compromise between the speed of operation and the system's complexity is encountered, as usual, whenever parallel or sequential machine operation is considered. The sequential methods (decision trees and "reward and punishment"* methods fall in this category) are often relatively slow in learning,

*The Perceptron is an example of a "reward and punishment" method in which a human observer rewards the machine by increasing the influence of those of the machine's parts that partake in a correct decision (as judged by the observer) and in which the machine is punished by reducing the influence of parts that contribute to incorrect decisions.

but they usually require less equipment than methods of classification that use parallel data-processing schemes. Although of a less fundamental nature than the matter of fixed or undetermined error criterion, the method of processing input information is of practical interest. It is probably not possible to decide categorically, once and for all, which of the two is better; they need to be judged and compared only if the common grounds for comparison are established. Such common grounds would be provided by evaluating the performance of two machines (one sequential and one parallel) that use the same error criterion and that try to solve the same problem. Only under such conditions does it seem reasonable to pass judgment on the merits of one method or another.

Most of the methods of pattern recognition require a certain amount of preprocessing before the application of the main technique. Figures are centered, rotated, normalized in size, lines are thinned out, gaps are closed, and so on. These operations are performed with the tacit assumption that category membership is unaffected by them. In general, this is a dangerous assumption to make; it is equivalent to making a partial statement of what constitutes similarity – a partial statement of the error criterion with which members of a category are judged to be close. It is readily appreciated that from a strictly operational point of view it is sometimes necessary to introduce some preliminary operations such as those mentioned above. In this case, however, the effect of the preliminary operations should be included in some other way to assure that the transformation from the stimulus (the figure, word, or other thing) to its representation is a one-to-one transformation. If this is not assured and, for example, figures are centered before the application of a technique, the possibility that the category definition should contain information about the position of the figure in the visual field is excluded. The coordinate shifts necessary to center the figure can be introduced as additional dimensions in the vector representation of a figure. If position in the visual field is indeed unimportant in deciding membership in a category, examples of the category will have a large variance in their position coordinates. This fact will be exhibited by the techniques outlined in this report, for the optimum transformation will deemphasize the position-indicating dimensions of the signal space. But if position in the visual field is of importance, the stimuli will be clustered along their position coordinates, a fact that the optimum transformation will certainly exhibit.

It is important to realize that special-purpose devices, in which the designer decides that certain properties of the input are not relevant to category membership, can be constructed, and substantial savings in instrumentation may be realized. A general theory and a general learning machine, on the other hand, should not allow representations of the input that destroy the one-to-one nature of the representation.

Recognition of membership in a category learned from a set of its known examples has many practical applications. Some of these will now be described without attempting to catalog them exhaustively.

In one set of applications, recognition of membership in a category is the desired end result. Recognition of a specific spoken word or speech sound independently of the talker who utters it is an application that belongs to this set. Samples of the word spoken by different talkers form the set of known examples of the category, and the meaning of the spoken word is the name of the category. Recognition of a speaker and his separation from other speakers independently of the words that are spoken is another practical application of the above type. The examples from which speaker recognition is learned are segments of speech by the desired speaker in which each segment is represented by a single vector. Recognition of a foreign accent and recognition of a foreign language are also problems of this type.

Recognition of a disease from all the symptoms of the patient is a further application of the above type. In this problem where the techniques described in this report are applied as automatic diagnostic aids, the patient's state, as measured by his symptoms, is represented as a vector in a multidimensional space. As stated earlier, the dimensions of the space are quantitative tests that can be performed on the patient. The patient's temperature, white blood cell count, sugar content as measured by urinanalysis, and the results of other laboratory chemical or bacteriological tests illustrate some of the dimensions of the space in which the patient, from a medical point of view, can be characterized by a vector. In these dimensions the coordinate values are analog quantities. In some of the other dimensions that represent questions which a patient may be asked and which may be answered with "yes" or "no", the coordinate value is a binary variable. Does it hurt when the abdomen is pressed? Yes or no. Is the patient's throat red? Labeled examples from which recognition of a disease and its differentiation from other diseases may be learned are patients with known diseases. In addition to its use as a diagnostic aid, the technique described may serve as a research tool to shed light on the nature of a disease and point out the most significant symptoms (the eigenvectors) whose cause must be investigated.

In a second set of applications of the techniques described in this report the recognition of membership in a category is only an intermediate result. Speech compression is an example. Consider the example in which the verbatim transmission of the text of a spoken message is desired. Assume for the purpose of this example that the speech bandwidth is 4 kc, that the speaker talks at the rate of 120 words per minute, and that there are 6 letters in a word, on the average (including the space after each word). The message rate of 12 letters per second in the teletype code of 5 bits per letter could be transmitted at the same rate that the speaker is talking if 60 bits per second were sent. The usual 30-cps bandwidth that is assumed to be sufficient to achieve this rate results in the bandwidth compression of the original speech by a factor of 133. Using a phonetic alphabet would result in a further increase of this factor. Pattern recognition is used here to recognize the spoken words, phonemes, or other building blocks of speech, and translate them into a correctly spelled text. It is by no means easy to accomplish this task in practice; some fundamental problems, such as the segmentation of speech into words, must be solved first. Similar problems arise in the transmission and storage of pictorial information.

While reading this report many questions have probably occurred to the reader and many have remained unanswered. The field of pattern recognition is new and invites questions – some touching on highly controversial issues. Through continuing work and better understanding of the problems, it is hoped that the number of unanswered questions will be reduced – to be replaced by new ones. Basic questions such as "Why use a geometrical approach supported by decision theory?" may be raised. Other valid approaches have been suggested, and several are under investigation by other researchers. These have not been discussed in this report. Even within the geometrical approach a large body of work remains to be done – work that this author intends to pursue.

Acknowledgement

It gives me great pleasure to acknowledge my indebtedness to Professor Henry J. Zimmermann for his friendship and guidance during my stay in the Graduate School of the Massachusetts Institute of Technology, and to Professor Y. W. Lee for encouraging me in hours of need. I am grateful to the many individuals with whom I discussed my problem and who helped me with suggestions and constructive criticism.

I especially wish to thank Dr. Manuel V. Cerrillo, whose inspiring personality made my association with the Research Laboratory of Electronics, M. I. T. , an event never to be forgotten.

APPENDIX A

The Solution of Eigenvalue Problems

The frequency with which eigenvalue problems occur in classificatory analysis and the difficulty with which they are solved warrants the careful examination of the available methods of solution. The slowest link in the computations involved in the process of forming category membership measuring functions is the solution of eigenvalue problems. It takes a considerable length of time for even a fast computer to solve for the eigenvalues and vectors of a large matrix. This limits the speed with which the influence of a new event is felt on the recognition process. In order that machine learning be carried out in essentially "real time," it is necessary to search for a physical phenomenon or a natural process that is the solution of an eigenvalue problem. The natural phenomenon must have enough controllable parameters to allow the setting up of an arbitrary positive definite symmetric matrix. The objective of this appendix is to focus attention on the importance of finding such a natural phenomenon and to give an example that – although not completely general, as we shall see, nor as practical as some would like – does demonstrate the feasibility of solving eigenvalue problems very rapidly.

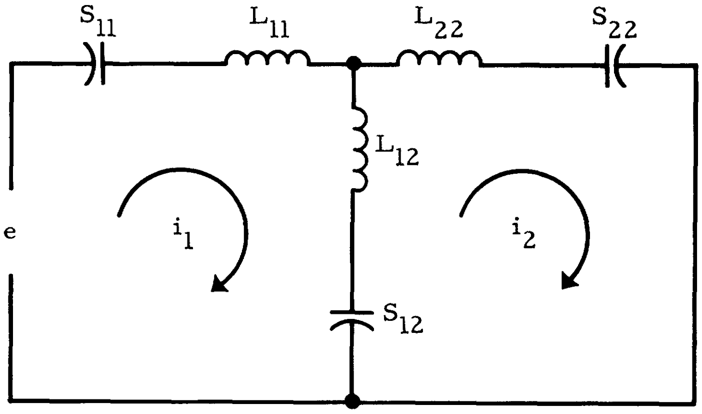


Fig. 15. Two-loop lossless network.

Consider the 2-loop lossless network of Fig. 15 that is excited with a voltage source, e, at its input. Letting the complex frequency be λ and the reciprocal capacitance (susceptance) values be S, we can write the loop equations of the network as

$$\begin{aligned} e_1 &= \left[\lambda(L_{11}+L_{12}) + \frac{S_{11}+S_{12}}{\lambda}\right] i_1 - \left[\lambda L_{12} + \frac{S_{12}}{\lambda}\right] i_2 \\ 0 &= -\left[\lambda L_{12} + \frac{S_{12}}{\lambda}\right] i_1 + \left[\lambda(L_{22}+L_{12}) + \frac{(S_{22}+S_{12})}{\lambda}\right] i_2 \end{aligned} \qquad (105)$$

Multiplying both sides of the equation by λ and writing it in matrix notation, we obtain Eq. 106, where e and i are, respectively, vectors of the voltage excitations in the loops and loop currents.

$$\lambda e = i(\lambda^2[L]+[S]) \tag{106}$$

The matrices [L] and [S] are

$$[L] = \begin{bmatrix} L_{11}+L_{12} & -L_{12} \\ -L_{12} & L_{22}+L_{12} \end{bmatrix} ; \; [S] = \begin{bmatrix} S_{11}+S_{12} & -S_{12} \\ -S_{12} & S_{22}+S_{12} \end{bmatrix} \tag{107}$$

If the input is short-circuited and the vector e is zero, any nonzero current that flows in the network must flow at frequencies that satisfy Eq. 108, where use is made of the knowledge that a lossless network must oscillate at pure imaginary frequencies λ = jω. The resulting equation is an eigenvalue problem of the same type encountered throughout this report.

$$[\lambda^2 L+C] = 0 = [C-\omega^2 L] \tag{108}$$

The matrix [L] is a completely arbitrary, symmetric, positive definite matrix in which the coefficients are each controlled by (at most) 2 circuit elements. The matrix [C] is also symmetric and positive definite, but its elements which are off the principal diagonal must be negative or zero. This does not have to be the case in the [L] matrix, for a negative mutual inductance is quite realizable. Note, however, that if the mutual capacitance is short-circuited, and the other capacitors are made equal (say, equal to one), then the natural frequencies of oscillation of the short-circuited network satisfy the eigenvalue problem of Eq. 109.

$$0 = \left[L - \frac{1}{\omega^2} I\right] \tag{109}$$

The most general 2-loop network represented by this equation is shown in Fig. 16, where a transformer replaces the mutual inductances.

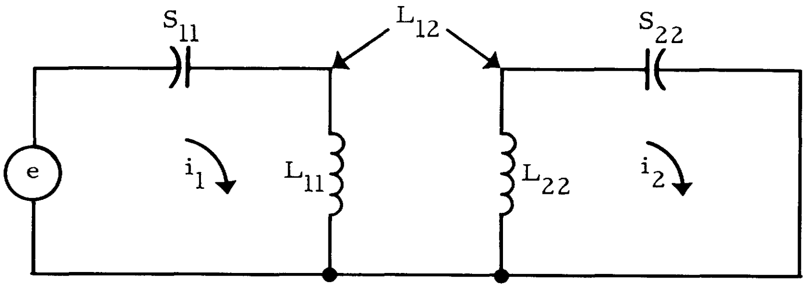


Fig. 16. Network solution of an eigenvalue problem.

The eigenvalues are the squares of the reciprocal natural frequencies of oscillation. Components of the eigenvector corresponding to a given eigenvalue are the magnitudes of the loop currents at the corresponding frequency.

Since lossless networks cannot be built, let us investigate the effect of losses in the network of Fig. 15. If a small resistance is connected in series with every inductance such that the frequency of oscillation is $\omega_d = \left(\omega_o^2 - \alpha^2\right)^{1/2}$, where α is the real part of the coordinates of the poles of the network, the error in using the damped natural frequencies ω_d in place of the undamped frequencies may be calculated. The percentage of error in determining the eigenvalues is given in Eq. 110, expressed in terms of the Q of the resonant circuits.

$$\text{Percentage of error in eigenvalues} = \frac{100}{(2Q)^2 - 1} \tag{110}$$

Even for a lossy network with a Q of 10, the error is only 0.25 per cent. We may thus draw the conclusion that network losses do not seriously affect the accuracy of the eigenvalues.

The eigenvalues may be obtained by spectrum analysis of any of the loop currents. This is readily accomplished by feeding the voltage across any of the series resistances into a tunable narrow-band filter whose tuning frequencies at peak outputs yield the eigenvalues. The corresponding eigenvector can be obtained by sampling the output amplitudes of synchronously tuned narrow filters connected to measure each loop current. The samples are taken when local peak outputs as a function of tuning are observed.

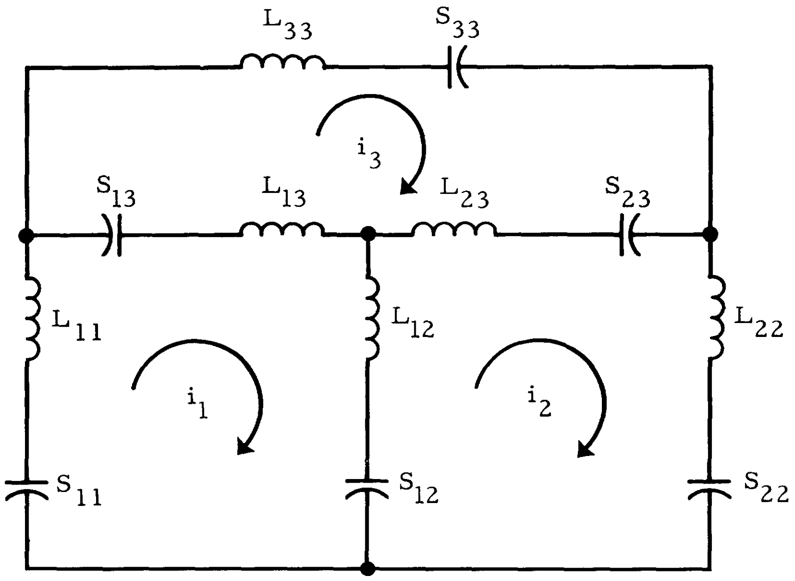


Fig. 17. Generalization of eigenvalue problem.

The size of the matrix solved by the preceding methods may be made arbitrarily large. The reader can readily verify that if the matrix whose eigenvalues and vectors we wish to compute is N × N, then the network topology has to consist of N nodes that are connected to each other and to ground by series LC networks as illustrated by Fig. 17 for N = 3.

APPENDIX B

Relationship Between the Measure of Similarity and the Likelihood Ratio

In this appendix the relationship between decisions based on a likelihood ratio and those made by the elementary version of the theory developed in Sections II and III will be discussed. We shall show that if the categories are statistically independent Gaussian processes with unknown but, in general, different means and variances, then the measure S of Section II measures contours of equal a posteriori probability. That is, the measure $S(x, \{F_m\})$, which measures the mean-square distance by a non-Euclidean metric between a point (vector) x and M members of an ensemble of points $\{F_m\}$, is a measure of the probability that x belongs to category F, given the observation x. Fixed values of S correspond to contours of equal a posteriori probability. The ratio of a posteriori probabilities is proportional to the likelihood ratio, the logarithm of which will be shown equal to the numerical output of the block diagram illustrated in Fig. 6.

Consider the situation in which an arbitrary event x may be a member of only one of two categories, F or G. The likelihood ratio that x belongs to F rather than to G is expressed by the ratio of a posteriori probabilities in Eq. 111. This equation may be simplified by Bayes' rule.

$$\frac{p(F/x)}{p(G/x)} = \frac{p(x/F)\ p(F)/p(x)}{p(x/G)\ p(G)/p(x)} \propto \frac{p_F(x)}{p_G(x)} = \ell(x) \tag{111}$$

The likelihood ratio is thus proportional to the ratio of the two joint probability densities of the two Gaussian processes. If membership in either of the two categories is equally likely, the proportionality becomes an equality.

Now let us examine the joint probability density $p_F(x)$, a factor of the likelihood ratio $\ell(x)$. For the multivariate Gaussian process the joint probability density is given by Eq. 112, where $[U]$ is the covariance matrix of F and $|U_{rs}|$ is the cofactor of the element with like subscripts in the covariance matrix. It should be noted that $|U_{rs}|/|U|$ is an element of U^{-1}.

$$p_F(x_1, x_2, \ldots, x_N) = \frac{1}{(2\pi)^{N/2}\ |U|^{1/2}} \exp\left[-\frac{1}{2}\sum_{r=1}^{N}\sum_{s=1}^{N}\frac{|U_{rs}|}{|U|}(x_r-m_r)(x_s-m_s)\right] \tag{112a}$$

$$= \frac{1}{(2\pi)^{N/2}\ |U|^{1/2}} \exp\left[-\frac{1}{2}\sum_{r=1}^{N}\sum_{s=1}^{N}\left[U_{rs}^{-1}\right](x_r-m_r)(x_s-m_s)\right] \tag{112b}$$

Contours of constant joint probability density occur for those values of x for which the argument of the exponential is constant. The exponent expressed in matrix notation is

$$\text{exponent} = \text{constant} = \left[-\frac{1}{2}(x-m_x)U^{-1}(x-m_x)^T\right] \tag{113}$$

We recall from the theorem of Section II that one of the operations on the set of points $\{F_m\}$ which the optimum metric performed was a rotation expressible by an orthogonal matrix $[C]$. This is a pure rotation (an orthonormal transformation), where columns of $[C]$ are unit eigenvectors of the covariance matrix U.

Let y be a new variable obtained from x by Eq. 114. Substituting Eq. 114b in Eq. 113, we obtain Eq. 115.

$$y = xC \tag{114a}$$

$$x = yC^{-1} \tag{114b}$$

$$\text{exponent} = \left[-\frac{1}{2}(y-m_y)C^{-1}U^{-1}[C^{-1}]^T(y-m_y)^T\right] \tag{115}$$

Since C is orthogonal, the special property of orthogonal matrices that $C^{-1} = C^T$ can be used to simplify Eq. 115. This yields

$$\text{exponent} = \left[-\frac{1}{2}(y-m_y)C^TU^{-1}C(y-m_y)^T\right] \tag{116}$$

Furthermore, since columns of C are eigenvectors of the covariance matrix U, the matrix C must satisfy Eq. 117a, where Λ is the diagonal matrix of eigenvalues of $[U-\lambda_nI] = 0$.

$$C^T[U-\Lambda]\, C = 0; \quad C^TUC = C^T\Lambda C = \Lambda \tag{117a}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{bmatrix} \tag{117b}$$

By taking the inverse of both sides of Eq. 117a and again employing the special property of orthogonal matrices, Eq. 118 may be obtained. This expression, when substituted in Eq. 116, produces the result stated in Eq. 119.

$$C^TU^{-1}C = \Lambda^{-1} \tag{118}$$

$$\text{exponent} = \text{constant} = \left[-\frac{1}{2}(y-m_y)\Lambda^{-1}(y-m_y)^T\right] \tag{119}$$

The quadratic form of Eq. 119 expresses the fact that contours of constant probability density are ellipsoids with centers at m_y, the direction of the principal axes is along eigenvectors of the covariance matrix, and the diameters are equal to the corresponding eigenvalues. Converting the quadratic form of Eq. 119 to a sum in Eq. 120 exhibits this result in a more familiar form, where y_n is the coordinate of y in the direction of the n^{th} eigenvector and m_n is the mean of the ensemble in the same direction.

$$\text{exponent} = \left[-\frac{1}{2}\sum_{n=1}^{N}\frac{(y_n-m_n)^2}{\lambda_n}\right] \tag{120}$$

An expression of identical appearance can be derived from the exponent of the joint probability density of category G. The differences between the two exponents are in the directions of their eigenvectors and the numerical magnitudes of their eigenvalues and ensemble means. Denoting the exponents in the two probability densities by f(x) and g(x), we can write the logarithm of the likelihood ratio as in Eq. 121, where K is a constant that involves the ratio of a priori probabilities and the ratio of determinants of the two covariance matrices.

$$\log \ell(x) = K + f(x) - g(x) \tag{121}$$

Now we shall show that $S(x, \{F_m\})$ = constant also satisfies Eq. 119, and that the decision, comparison of Eq. 122 with a constant, is identical with basing decisions on the likelihood ratio by means of Eq. 121. It will be recalled that S is the mean-square Euclidean distance between x and members of $\{F_m\}$ after both are transformed by a linear transformation that consists of a rotation [C] and a diagonal transformation [W]. The rotation [C] is identical with that defined earlier in this appendix, and the elements of the diagonal transformation are the reciprocal standard deviations of $\{F_m\}$ in the direction of the eigenvectors. The mean-square distance S may be expressed by Eq. 122. With the use of Eq. 114b this can be simplified to Eq. 123, where Y_m is the transformation of F_m.

$$S(x, \{F_m\}) = \overline{(x-F_m)\, CWIW^T C^T (x-F_m)^T}^{\,m} = \text{constant} \tag{122}$$

$$= \overline{(y-Y_m)\, WIW^T (y-Y_m)^T}^{\,m} \tag{123}$$

But WIW^T is a diagonal matrix with elements equal to the reciprocal variances of $\{F_m\}$. From a comparison of Eqs. 24c and 40 we see that the variances are equal to the eigenvalues; and we obtain Eq. 124. Substituting this in Eq. 123 yields Eq. 125.

$$WIW^T = \Lambda^{-1} \tag{124}$$

$$\text{constant} = \overline{(y-Y_m)\, \Lambda^{-1} (y-Y_m)^T}^{\,m} \tag{125}$$

Equation 125 can be brought into the form of the exponent of the joint probability density, Eq. 120. Writing Eq. 125 as a sum and bringing averaging under the summation sign yield

$$\text{constant} = \overline{\sum_{n=1}^{N} \frac{(y_n - Y_{mn})^2}{\lambda_n}}^{m} = \sum_{n=1}^{N} \left[\overline{\frac{(y_n - Y_{mn})^2}{\lambda_n}}^{m} \right] \tag{126}$$

Expanding the square and adding and subtracting $\overline{Y_n}^2$ from each term of the numerator result in Eq. 127.

$$\text{constant} = \sum_{n=1}^{N} \frac{y_n^2 - 2y_n\overline{Y_n} + \overline{Y_n^2}}{\lambda_n} = \sum_{n=1}^{N} \frac{y_n^2 - 2y_n\overline{Y_n} + \overline{Y_n}^2 + \overline{Y_n^2} - \overline{Y_n}^2}{\lambda_n} \tag{127}$$

Since $\overline{Y_n} = m_n$, and $\overline{Y_n^2} - \overline{Y_n}^2 = \sigma_n^2 = \lambda_n$, this equation can be written in the simplified form of Eq. 128, which is recognized as containing the exponent f(x) of the joint probability density.

$$\text{constant} = \sum_{n=1}^{N} \frac{(y_n - m_n)^2 + \lambda_n}{\lambda_n} = N + \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n} = N + f(x) \tag{128}$$

The difference of the two measures – the mean-square distance of x to category F minus its mean-square distance to category G – is given in Eq. 129 and is seen to be a measure of the logarithm of the likelihood ratio expressed in Eq. 121.

$$\text{threshold } T \gtrless S_F(x, \{F_m\}) - S_G(x, \{G_m\}) \tag{129a}$$

$$T \gtrless f(x) - g(x) \tag{129b}$$

APPENDIX C

Automatic Recognition of Spoken Words

In this appendix we would like to illustrate the application of the theory to a practical problem in which membership in categories must be recognized by an automatic machine. The method discussed in Section II will be employed in the automatically "learned" recognition of speech events. These events will belong to one of ten different categories: the spoken words "zero", "one", "two", . . . , "nine". Each of these categories is represented by a number of utterances of the spoken digits made by a selection of male speakers. Four hundred different utterances by 10 male speakers with regional accents drawn from the northeast corner of the United States were used in this experiment. No other attempt was made to control the selection of speakers or their rate of speech. The choice of categories, indeed, even the selection of speech rather than of other types of events, was dictated in part by the availability of an automatic machine to convert speech to a set of vectors and in part by the desire to solve a practical problem.

Vector Representation of Spoken Words

The model of the physical world considered as adequate for representing the speech events was obtained through use of an 18-channel Vocoder.* The Vocoder consists of a set of 18 stagger-tuned bandpass filters and envelope detectors that produce, at the envelope detector outputs, the "instantaneous" frequency spectrum of the speech event as a function of time. A representative printout is shown in Fig. 18, where frequency is plotted vertically, time is plotted horizontally, and the intensity of the spectrum at a given frequency and time is proportional to the grey level of the sonograph recording at the corresponding time-frequency point. The numerical printout of Fig. 18 is obtained by digitizing the sonograph records into 18 frequency channels, each sampled at the rate of 20 msec/sample. Note that the samples are orthogonal by construction because they represent waveforms that are disjointed either in frequency or time. The resulting cell structure in the time-frequency plane represents a 1-sec duration speech event as a vector in a 900-dimensional space. Each dimension corresponds to a possible cell location, and the coordinate value of a dimension equals the intensity of the corresponding cell. In Fig. 18, a 3-digit binary number (8 levels) represents the sonograph intensity, after the instantaneous total speech intensity has been normalized by the action of a fast AGC. To increase the resemblance of the printout in the lower portion of Fig. 18 to the sonograph of the upper portion, the grey level 1 was suppressed.

This model of the physical world is known to be more than adequate in containing category defining information; good quality intelligible speech can be synthesized from

*The Vocoder used was the property of the Air Force Cambridge Research Center and was made available through the courtesy of C. P. Smith.

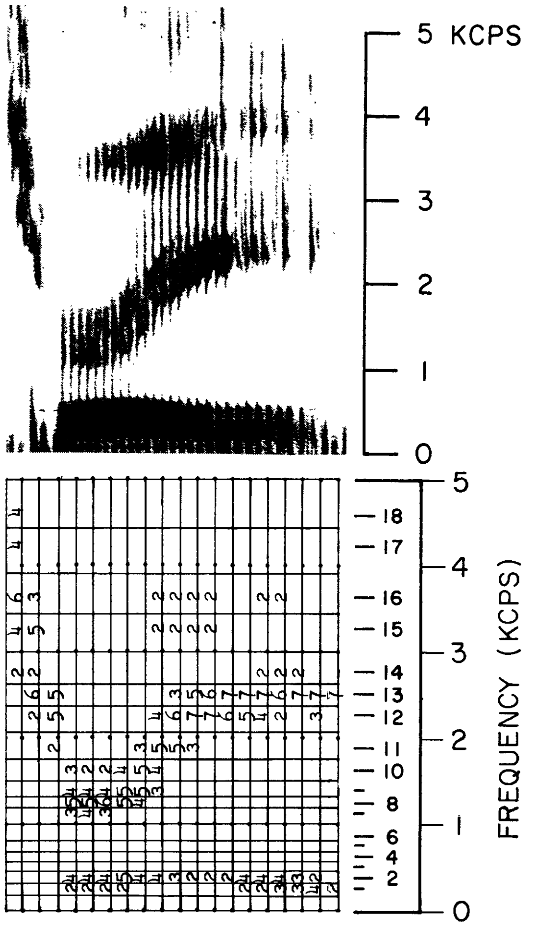


Fig. 18. Sonograph representation of the spoken word "three."

the previously described characterization of speech.

It should be noted that many alternative representations of speech events are possible. Direct samples of bandlimited speech waveforms sampled at the Nyquist rate, or Fourier series expansion coefficients would allow the representation of a speech event as a vector in a high-dimensional space. The number of necessary dimensions would be somewhat higher than 900, however. An additional advantage of the quantized sonograph representation of speech events is that in the signal space constructed from it, like events would exhibit a fair degree of clustering, a fact that facilitates the demonstration of the special linear theory of Section II.

The spoken digits are presented in a sentence, where parts of other words bracket the digit of interest to us. In this particular example the beginning and end of a word are readily identified, however, by the short silent intervals at either extreme. It is assumed in the following discussion that the beginning and end of a word have been found either by some auxiliary method or because words are spoken in isolation; and the recording tape speed, or sampling rate, has been adjusted so that an equal number of samples of each word is taken. Although instrumented with comparative ease, it is not at all necessary to assume such a simple normalization of speech events. Perhaps more realistic and more practical normalizations could be formulated to account for the varying word lengths that result from the differences in the number of syllables of words and the differences in the speaker speed. Since the purpose of this experiment was not to solve a particular problem in speech recognition but to illustrate the analysis technique developed, the simple normalization is assumed.

Each spoken digit is represented by a 361-dimensional vector in a vector space of 361 dimensions. This vector space is constructed by taking 20 equally spaced samples per word of each of the 18 Vocoder channel outputs. The resulting 360 dimensions are augmented by the word length considered as an additional dimension. By retaining word-length information, the one-to-one mapping of spoken words into a fixed number of dimensional vectors is achieved. The change of sample spacing between real time words and those in normalized time made it necessary to interpolate between the heights of adjacent samples in the same Vocoder channel in order to obtain the sample heights that would correspond to the new sample spacing. Linear interpolation was employed.

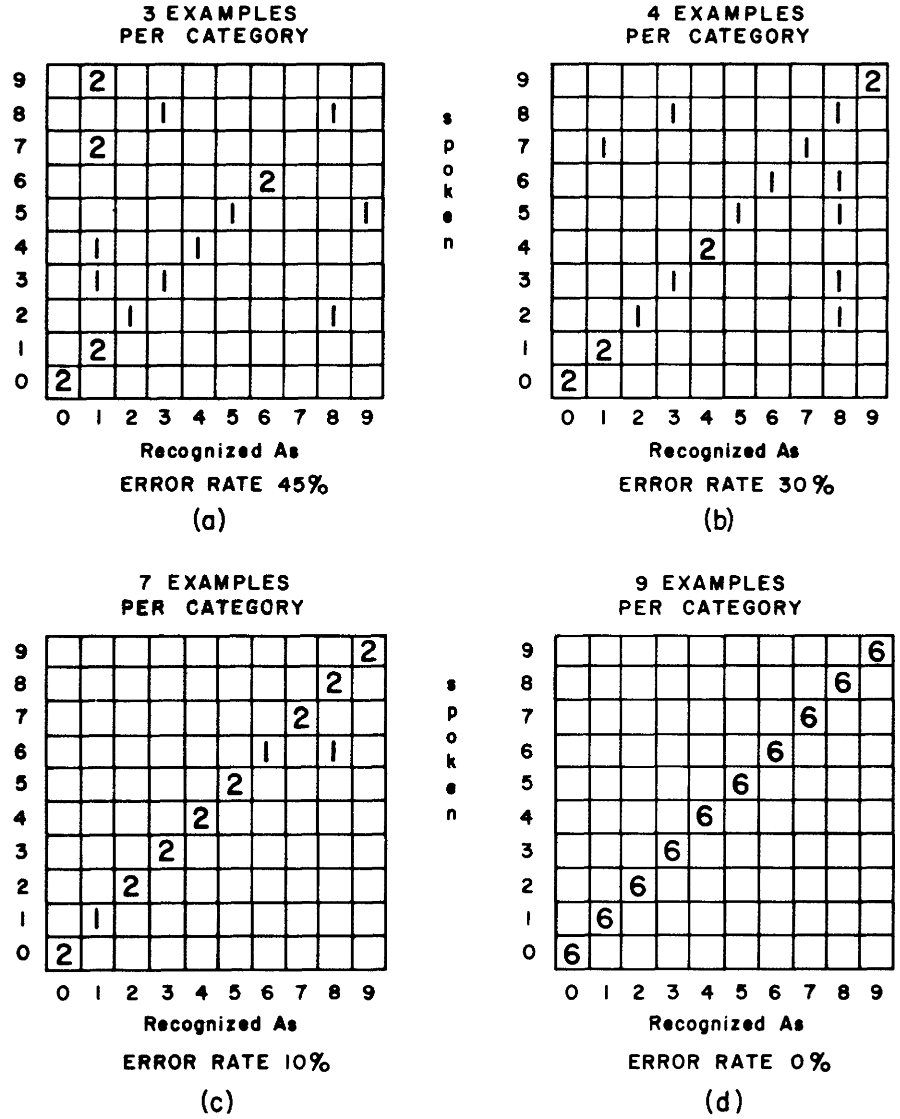
Computation of Optimum Transformations

First, M samples of each of the 10 categories were selected as the labeled samples from which numeral recognition is learned. At first M was chosen as 3; later the computations were repeated with M gradually increasing to 10.

The first step in the process of finding the orthogonal transformations that minimize the mean-square distances between points of each of the categories is to find the orthonormal transformations that rotate the given coordinate system into the optimum ones. Because this would involve solving for the eigenvalues and vectors of 361 X 361 matrices, a time-consuming process, we utilize the knowledge that the eigenvectors of each solution will be contained in the M-dimensional linear manifold of the space spanned by the vectors of each set. The M vectors of each set were therefore orthogonalized to obtain 10 M-dimensional coordinate systems in which the sample vectors of each set could be expressed in no more than M-dimensions.

The covariance matrices of each set of vectors were constructed, and the eigenvalues and eigenvectors of these matrices were obtained. All of these computations were performed on a (RECOMP II) general-purpose digital computer. The eigenvectors of a covariance matrix form the columns of the rotation transformation C of Section II. The reciprocals of the corresponding eigenvalues are the elements of the diagonal matrix W, which expresses the weighting associated with each new eigenvector. The eigenvectors were then expressed in the original 361-dimensional coordinate system, and the computation of the quadratic forms according to Eq. 128 was programmed on the computer. The decision regarding membership of an unknown speech event was determined by the decision rule of Eq. 129 by choosing the category corresponding to the quadratic form of smallest value.

Typical results which demonstrate improvement in the machine's performance as the



3 EXAMPLES PER CATEGORY

	0	1	2	3	4	5	6	7	8	9
9		2								
8				1					1	
7		2								
6							2			
5						1				1
4		1			1					
3		1		1						
2			1						1	
1		2								
0	2									

Recognized As

ERROR RATE 45%

(a)

4 EXAMPLES PER CATEGORY

	0	1	2	3	4	5	6	7	8	9
9										2
8				1					1	
7		1						1		
6							1		1	
5						1			1	
4					2					
3				1					1	
2			1						1	
1		2								
0	2									

Recognized As

ERROR RATE 30%

(b)

7 EXAMPLES PER CATEGORY

	0	1	2	3	4	5	6	7	8	9
9										2
8									2	
7								2		
6							1		1	
5						2				
4					2					
3				2						
2			2							
1		1								
0	2									

Recognized As

ERROR RATE 10%

(c)

9 EXAMPLES PER CATEGORY

	0	1	2	3	4	5	6	7	8	9
9										6
8									6	
7								6		
6							6			
5						6				
4					6					
3				6						
2			6							
1		6								
0	6									

Recognized As

ERROR RATE 0%

(d)

Fig. 19. Confusion matrices illustrating the process of learning numeral recognition.

number of known, labeled examples of spoken digits is increased, are illustrated in Fig. 19. This figure contains 4 confusion matrices constructed for the cases in which numeral recognition was learned from 3, 4, 7, and 9 examples of each of the 10 categories of digits. The ordinate of a cell in the matrix signifies the digit that is spoken, the abscissa denotes the decision of the machine, and the number in the cell states the number of instances in which the stated decision was made. The number 1 in row 6 and column 8, for example, denotes the fact that in one instance a spoken digit 6 was recognized as an 8. Note that the error rate decreases as the number of known examples of categories is increased. For 9 examples per category no errors were made. This result is particularly interesting in view of the fact that many of the spoken digits that were tested were spoken by persons not included among those whose words were used as examples.

Using the same techniques, but enlarging the number of dimensions of the space in which spoken digits are represented, we may obtain improved results. The addition of parameters – considered useful from the standpoint of linguistics – as new dimensions would increase the clusterability of words of the same category. An example of such a parameter is "the number of intervals of time into which a word may be divided so that an interval is defined in a special way." The interval is a segment of time during which the mean-square difference of successive "instantaneous" spectra is less than a suitably chosen threshold value. The number of intervals defined in this manner is probably indicative of the number of different sounds that exist in a spoken digit. The number of different sounds per digit is expected to be substantially the same for a given digit regardless of who speaks it, but this number may differ for another spoken digit.

APPENDIX D

General Equivalence of the Likelihood Ratio and the Geometrical Decision Rule

It will be shown in this appendix that the decision rule derived from geometrical considerations in which members of sets are clustered while sets are kept apart is equivalent to a Bayes' rule under very general assumptions regarding the probability densities associated with the distribution of members of sets.

The decision rule stated in Section V is to:

$$\begin{aligned} &\text{decide } v \in F \text{ if } \left|u(v)-\overline{u_F(v)}\right| < \left|u(v)-\overline{u_G(v)}\right| \\ &\text{decide } v \in G \text{ if } \left|u(v)-\overline{u_F(v)}\right| > \left|u(v)-\overline{u_G(v)}\right| \end{aligned} \tag{130}$$

This decision rule compares the value of the surface over the point to be classified, $u(v)$, with the average values of the surface over the two sets of given samples, $\overline{u_F(v)}$ and $\overline{u_G(v)}$. The decision is made that v is a member of the set, F or G, to which it is closer.

If $p_F(v)$ and $p_G(v)$ are the probability densities of v under the assumptions that v is a member of class F or G, then by Bayes' rule we have

$$\begin{aligned} &\text{decide } v \in F \text{ if } p_F(v) > p_G(v) \\ &\text{decide } v \in G \text{ if } p_F(v) < p_G(v) \end{aligned} \tag{131}$$

This decision rule, if the a priori probabilities of F and G are equal, calls for deciding that v is a member of F if, under the assumption that F occurred, v is a more likely observation than under the assumption that G occurred.

It will be shown in this appendix that if the function $u(v)$ involved in Eq. 130 is obtained as a solution of the minimization problem stated in Section V and given below in Eqs. 132 and 133, then the two decision rules, Eqs. 130 and 131, are equivalent. The equivalence is true, of course, only as the sample size approaches infinity.

$$\text{minimize} \left[\sigma_F^2(u) + \sigma_G^2(u)\right] = Q \tag{132}$$

if

$$\bar{u}_F - \bar{u}_G = K > 0 \tag{133}$$

The significance of this minimization problem is that it obtains the function u required in Eq. 130 by operating on a set of samples. The solution of the problem under the assumption of a certain class of polynomial transformations was given in Section V. Here we shall show that the solution equates the decision rules of Eqs. 130 and 131 under the assumption of quite arbitrary classes of transformations.

First, the rule of Eq. 130 will be reduced to an equivalent but simpler form. Note that if the inequalities of Eq. 130 hold, so do those of Eq. 134.

$$\begin{aligned} &\text{decide } v \in F \text{ if } [u(v)-\bar{u}_F]^2 < [u(v)-\bar{u}_G]^2 \\ &\text{decide } v \in G \text{ if } [u(v)-\bar{u}_F]^2 > [u(v)-\bar{u}_G]^2 \end{aligned} \qquad (134)$$

Squaring both sides of the inequalities, cancelling like terms, and rearranging the inequality yield Eq. 135.

$$u^2(v) - 2u\bar{u}_F + \bar{u}_F^2 < u^2(v) - 2u\bar{u}_G + \bar{u}_G^2 \qquad (135a)$$

$$\bar{u}_F^2 - \bar{u}_G^2 < 2u(\overline{u_F}-\overline{u_G}) \qquad (135b)$$

$$\frac{\bar{u}_F + \bar{u}_G}{2} < u(v) \qquad (135c)$$

Now we solve the minimization problem of Eq. 132, subject to the constraint stated in Eq. 133. To facilitate the solution, we minimize Eq. 136a subject to Eq. 136b instead, and show later that the two problems have identical solutions.

$$\text{minimize } Q' = \overline{u_F^2} + \overline{u_G^2} \qquad (136a)$$

$$K = \overline{u_F} - \overline{u_G} > 0 \qquad (136b)$$

Using the method of Lagrange multipliers and writing out the expression in integral form, we obtain Eq. 137.

$$Q' + \lambda K = \int \left\{ u^2(v)[p_F(v)+p_G(v)] + \lambda u(v)[p_F(v)-p_G(v)] \right\} dv \qquad (137)$$

Setting the variation of Eq. 137 equal to zero, Eq. 138 is obtained.

$$\delta(Q'+\lambda K) = 0 = \int \{2u(v)[p_F(v)+p_G(v)] + \lambda[p_F(v)-p_G(v)]\} \, \delta u \, dv \qquad (138)$$

Since δu is arbitrary, however, the expression { } must be identically zero for all δu. Solving for u(v), we obtain Eq. 139.

$$u(v) = -\frac{\lambda}{2} \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} \qquad (139)$$

The Lagrange multiplier may now be evaluated by substituting Eq. 139 in the constraint of Eq. 136b.

$$K = \bar{u}_F - \bar{u}_G = -\frac{\lambda}{2} \int \frac{[p_F(v)-p_G(v)]^2}{[p_F(v)+p_G(v)]} dv > 0 \qquad (140)$$

We have to note only that the integral is a positive quantity, establishing the fact that the constant λ is negative. This fact permits us to rewrite Eq. 139 as

$$u(v) = \left|\frac{\lambda}{2}\right| \left(\frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} \right) \tag{141}$$

Substitution in the simplified decision rule, Eq. 135c, yields the further simplified rule of Eq. 142.

$$u(v) > \frac{\bar{u}_F + \bar{u}_G}{2} = \left|\frac{\lambda}{4}\right| \int [p_F(v)-p_G(v)]\, dv = 0 \tag{142a}$$

$$\text{decide } v \in F \text{ if } u(v) > 0 \tag{142b}$$

$$\text{decide } v \in G \text{ if } u(v) < 0 \tag{142c}$$

When Eq. 141 is substituted in the decision rule of Eq. 142, we obtain Eq. 143, which is satisfied only if $p_F(v) > p_G(v)$, the criterion required by Bayes' rule.

$$\text{decide } v \in F \text{ if } u(v) = \left|\frac{\lambda}{2}\right| \frac{p_F(v) - p_G(v)}{p_F(v) + p_G(v)} > 0 \tag{143}$$

Thus the two rules are proved to be equivalent if u(v) is a solution of the minimization problem stated in Eqs. 136a and 136b. It now remains to show only that the problem stated in Eqs. 134 and 133 is equivalent to that stated in Eqs. 136a and 136b.

It is seen from Eqs. 142a and 133 that $\overline{u_F(v)} = -\overline{u_G(v)} = K/2$. Substituting this value in Eq. 132, we obtain Eq. 144.

$$Q = \sigma_F^2(u) + \sigma_G^2(u) = \overline{u_F^2} + \overline{u_G^2} - \overline{u_F}^2 - \bar{u}_G^2 = \overline{u_F^2} + \overline{u_G^2} - \frac{K^2}{2} \tag{144}$$

Since $K^2/2$ is a constant, the minimization of Q leads to the same solution as the minimization of Q'.

Bibliography

1. Ashby, W. Ross. Design for a Brain (John Wiley and Sons, Inc., New York and Chapman and Hall, Ltd., London, 1952).

2. Bar-Hillel, Y. Can translation be mechanized? Am. Scientist 42, 248-260 (April 1954).

3. Bar-Hillel, Y. Linguistic problems connected with machine translation, Brit. J. Philos. Sci. 20, No. 3, 217-225 (July 1953).

4. Bar-Hillel, Y. The present state of research on mechanical translation, Am. Document. 2, 229-237 (1952).

5. Bernstein, A., et al. A chess playing program for the IBM 704, Proc. Western Joint Computer Conference, May 6-8, 1958 (AIEE, New York), pp. 157-159.

6. Bomba, J. S. Alpha-numeric character recognition using local operations, Proc. Eastern Joint Computer Conference, December 3, 1959 (IRE, New York), pp. 218-224.

7. Bremer, R. W. A checklist of intelligence for programming systems, Communs. Assoc. Computing Machinery 2, 8-13 (March 1959).

8. Carr, J. W. Recursive subscripting compilers and list-type memories, Communs. Assoc. Computing Machinery 2, 4-6 (February 1959).

9. Chow, C. K. An optimum character recognition system using decision functions, Trans. IRE, EC-6, No. 4, 247-254 (December 1957).

10. Clark, W. A., and Farley, B. G. Generalization of pattern recognition in a self-organizing system, Proc. Western Joint Computer Conference, March 1, 1955 (IRE, New York), pp. 86-91.

11. David, E. E., Jr. Artificial auditory recognition in telephony, IBM J. Res. Develop. 2, No. 4, 294-301 (1958).

12. David, E. E., Jr., and McDonald, H. S. A bit-squeezing technique applied to speech signals, IRE Convention Record, Part 4, 1956, pp. 148-152.

13. Denes, P. The design and operation of the mechanical speech recognizer, J. Brit. I. R. E. 19, 219-229; Discussion, 230-234 (April 1959).

14. Dimond, T. L. Devices for reading handwritten characters, Proc. Eastern Joint Computer Conference, 1957, pp. 232-237.

15. Dinneen, G. P. Programming pattern recognition, Proc. Western Joint Computer Conference, March 1955.

16. Dunker, K. On Problem Solving, Psychol. Monogr., Vol. 58, No. 270 (1945).

17. Evey, R. J. Use of a computer to design character recognition logic, Proc. Eastern Joint Computer Conference, December 3, 1959, pp. 205-211.

18. Feldman, J. A theory of binary choice behavior, CIP Working Paper No. 12, Carnegie Institute of Technology, Pittsburgh, May 1958.

19. Flores, I. An optimum character recognition system using decision functions, Trans. IRE, EC-7, 180 (June 1958).

20. Friedberg, R. M. A learning machine, Part I, IBM J. Res. Develop. 2, 2-13 (January 1958).

21. Fucks, W. On mathematical analysis of style, Biometrika 39, 122 (1952).

22. Galanter, Eugene H. The behavior of thought, Paper presented at the American Psychological Association Meeting, Chicago, 1956.

23. Gardner, M. Logic Machines and Diagrams (McGraw-Hill Book Company, Inc., New York, 1958).

24. Gelernter, H. L., and Rochester, N. Intelligent behavior in problem-solving machines, IBM J. Res. Develop. 2, 336-345 (October 1958).

25. Gentzen, Gerhard. Untersuchungen über das logische Schliessen, Math. Z. 39, 176-210; 405-431 (1934).

26. Glantz, H. T. On the recognition of information with a digital computer, J. Assoc. Computing Machinery (April 1957).

27. Gold, B. Machine recognition of hand-sent Morse code, Trans. IRE, IT-5, 17-24 (March 1950).

28. Greanias, E. C., et al. Design of logics for recognition of printed characters by simulation, IBM J. Res. Develop. 1, 8-18, January 1957.

29. Grimsdale, R. L., et al. A system for the automatic recognition of patterns, J. Inst. Elec. Engrs. (London) 106, Part B, 210-221 (March 1959).

30. Harris, Robert T., and Jarrett, J. L. Language and Informal Logic (Longmans Green and Co., Inc., New York, 1956).

31. Hebb, D. O. The Organization of Behavior (John Wiley and Sons, Inc., New York, and Chapman and Hall, Ltd., London, 1949).

32. Hilgard, E. Theories of Learning, Second Ed. (Appleton-Century-Crofts, Inc., New York, 1956).

33. Hovland, C. I. A "communication analysis" of concept learning, Psychol. Rev. 59, 461-472 (1952).

34. Humphrey, G. Thinking (John Wiley and Sons, Inc., New York, 1951).

35. Ianov, Yu. I. On equivalency and transformations of program schemes, Doklady Akad. Nauk S. S. S. R. 113, No. 1, 39-42 (1957).

36. Ianov, Yu. I. On matrix program schemes, Doklady Akad. Nauk S. S. S. R. 113, No. 2, 283-286 (1957); also published in Communs. Assoc. Computing Machinery 1, No. 12 (December 1958).

37. Kirsch, R. A., and others. Experiments in processing pictorial information with a digital computer, Proc. Eastern Joint Computer Conference, December 3, 1957, pp. 221-230.

38. Kister, J., and others. Experiments in chess, J. Assoc. Computing Machinery 4, 2 (April 1957).

39. Kramer, H. P., and Mathews, M. V. A linear coding for transmitting a set of correlated signals, Trans. IRE, IT-2 (September 1956). (Paper presented at Symposium on Information Theory held at Massachusetts Institute of Technology, Cambridge, Mass., September 10-12, 1956.)

40. Kretzmer, E. R. Reduced alphabet representation of television signals, IRE Convention Record, Part 4, 1956, p. 40.

41. Lambek, J. The mathematics of sentence structure, Amer. Math. Monthly 65, 154-170 (1958).

42. Lashley, K. S. Cerebral Mechanism in Behavior (John Wiley and Sons, Inc., New York, 1951).

43. Latil, P. de. Thinking by Machine (Houghton-Mifflin Co., Boston, 1956).

44. Levin, K. Principles of Topological Psychology (McGraw-Hill Book Company, Inc., New York, 1936).

45. Locke, W. N., and Booth, A. D. Machine Translation of Languages (John Wiley and Sons, Inc., New York, 1955).

46. Luchins, A. S. Mechanization in problem solving, Psychol. Monogr. 54, No. 6 (1942).

47. Luhn, H. P. The automatic creation of literature abstracts, IBM J. Res. Develop. 2, 159-165 (April 1958).

48. Mattson, R. L. A self organizing logical system, paper presented at the Eastern Joint Computer Conference, December 3, 1959.

49. McCarthy, J. The inversion of functions defined by Turing machines, Automata Studies, edited by C. E. Shannon and J. McCarthy (Princeton University Press, Princeton, N. J., 1956), pp. 177-181.

50. McCulloch, W. S., and Pitts, W. H. A logical calculus of the ideas imminent in nervous activity, Bull. Math. Biophys. 9, 127 (1947).

51. McCulloch, W. S., and others. Symposium on the design of machines to simulate the behavior of the human brain, Trans. IRE, EC-5, No. 4 (December 1956).

52. The Mechanization of Thought Processes, Computer Bulletin 2, 92-93 (April-May 1959).

53. Miller, G. A. Language and Communication (McGraw-Hill Book Company, Inc., New York, 1951).

54. Miller, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information, Psychol. Rev. 63, 81-97 (1956).

55. Miller, G. A., and Selfridge, J. A. Verbal context and the recall of meaningful material, Am. J. Psychol. 63, 176-185 (1956).

56. Minsky, M. L. Exploration systems and syntactic processes, Summer Research Project on Artificial Intelligence, Dartmouth College, New Hamshire, 1956. (Unpublished report.)

57. Minsky, M. L. Heuristic aspects of the artificial intelligence problem, Group Report 34-35, Lincoln Laboratory, M. I. T., 17 December 1956, p. I-1-1-24.

58. Moore, O. K., and Anderson, S. B. Modern logic and tasks for experiments on problem solving behavior, J. Psychol. 38, 151-160 (1954).

59. More, T., Jr. Deductive Logic for Automata, S. M. Thesis, Department of Electrical Engineering, M. I. T., 1957.

60. Morris, C. Signs, Language and Behavior (Prentice-Hall, Inc., New York, 1946).

61. Neumann, J. von. The general and logical theory of automata, Cerebral Mechanism in Behavior, edited by W. Jeffress (John Wiley and Sons, Inc., New York, 1951).

62. Neumann, J. von. Theory of Games and Economic Behavior (Princeton University Press, Princeton, N. J., 1947).

63. Newell, A. The Chess Machine, Proc. Western Joint Computer Conference, March 1955.

64. Newell, A., Shaw, J. C., and Simon, H. A. Chess-playing programs and the problem of complexity, IBM J. Res. Develop. 2, No. 4, 320-335 (October 1958).

65. Newell, A., Shaw, J. C., and Simon, H. A. Elements of a Theory of Human Problem Solving, Report No. P-971, The Rand Corporation, Santa Monica, Calif., 4 March 1957.

66. Newell, A., Shaw, J. C., and Simon, H. A. The elements of a theory of human problem solving, Psychol. Rev., Vol. 65 (March 1958).

67. Newell, A., Shaw, J. C., and Simon, H. A. Empirical exploration of the logic theory machine, Proc. Western Joint Computer Conference, February 1957.

68. Newell, A., Shaw, J. C., and Simon, H. A. Empirical Exploration of the Logic Theory Machine (revised), Report No. P-951, The Rand Corporation, Santa Monica, Calif., March 14, 1957.

69. Newell, A., Shaw, J. C., and Simon, H. A. General Problem Solving Program, CIP Working Paper No. 7, Carnegie Institute of Technology, Pittsburgh, December 1957.

70. Newell, A., Shaw, J. C., and Simon, H. A. The Processes of Creative Thinking, Report No. P-1320, The Rand Corporation, Santa Monica, Calif., August 1958.

71. Newell, A., Shaw, J. C., and Simon, H. A. Report on a General Problem Solving Program, Report No. P-1584, The Rand Corporation, Santa Monica, Calif., January 1959.

72. Newell, A., and Shaw, J. C. Programming the logic theory machine, Proc. Western Joint Computer Conference, February 1957.

73. Newell, A., and Shaw, J. C. Programming the Logic Theory Machine (revised), Report No. P-954, The Rand Corporation, Santa Monica, Calif., 28 February 1957.

74. Newell, A., and Simon, H. A. Current Developments in Complex Information Processing, Report No. P-850, The Rand Corporation, Santa Monica, Calif., May 1, 1956.

75. Newell, A., and Simon, H. A. The logic theory machine, Trans. IRE, IT-2, No. 2, 61-79 (September 1956).

76. Newell, A., and Simon, H. A. The Logic Theory Machine. A Complex Information Processing System (revised), Report No. P-868, The Rand Corporation, Santa Monica, Calif., 12 July 1956.

77. Oettinger, A. G. Simple Learning by a Digital Computer, Proc. Assoc. Computing Machinery, September 1952.

78. Perry, J. W., Kent, A., and Berry, N. M. Machine Literature Searching (Interscience Publishers, Inc., New York, 1956).

79. Pitts, W. H., and McCulloch, W. S. How we know universals, the perception of auditory and visual form, Bull. Math. Biophys. 9, 1048 (1947).

80. Pólya, G. How to Solve It (Princeton University Press, Princeton, N. J., 1945).

81. Pólya, G. Mathematics and Plausible Reasoning, Vols. I and II (Princeton University Press, Princeton, N. J., 1954).

82. Rapaport, D. The Organization and Pathology of Thought (Columbia University Press, New York, 1951).

83. Rochester, N., and others. Tests on a Cell Assembly Theory of the Action of the Brain Using a Large Digital Computer, Trans. IRE, IT-2, No. 3, September 1956.

84. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65, 6 (1958).

85. Rosenblatt, F. The Perceptron, A Theory of Statistical Separability in Cognitive Systems, Cornell Aeronautical Laboratory, Project PARA, Report No. VG-1196-G-1 (January 1958).

86. Selfridge, O. G. Pandemonium: a paradigm for learning, Proceedings of the Symposium on Mechanization of Thought Processes, National Physical Laboratory, Teddington, Middlesex, England, November 24-27, 1958.

87. Selfridge, O. G. Pattern recognition and learning, Symposium on Information Theory, London, England (1955). Preprinted Group Report 34-43, Lincoln Laboratory of Massachusetts Institute of Technology, July 20, 1955.

88. Selfridge, O. G. Pattern recognition and modern computers, Proc. Western Joint Computer Conference, pp. 91-93 (March 1955).

89. Selfridge, O. G., et al. Pattern recognition and reading by machine, Proc. Eastern Joint Computer Conference, December 3, 1959.

90. Shannon, C. E. Communication theory of secrecy systems, Bell System Tech. J. 28, 656-715 (1949).

91. Shannon, C. E. Computers and automata, Proc. IRE 41, 1235-1241 (1953).

92. Shannon, C. E. A mathematical theory of communication, Bell System Tech. J. 27, 379-423 (1948).

93. Shannon, C. E. Prediction and entropy of printed English, Bell System Tech. J. 30, 50-64 (1951).

94. Shannon, C. E. Programming a computer for playing chess, Phil. Mag. 41, 256-275 (1950).

95. Shannon, C. E. The rate of approach to ideal coding, IRE National Convention Record, Part 4, 1955 (Abstract pages only).

96. Shannon, C. E. A universal Turing machine with two internal states, Automata Studies, edited by C. E. Shannon and J. McCarthy, Annals of Mathematics Studies No. 34 (Princeton University Press, Princeton, N. J., 1956), pp. 157-165.

97. Shaw, J. C., and others. A command structure for complex information processing, Proc. Western Joint Computer Conference, May 1958.

98. Simon, H. A. A behavioral model of rational choice, Quart. J. Econ. 69, 99-118 (1955).

99. Simon, H. A. Rational choice and the structure of the environment, Psychol. Rev. 63, 129-138 (1956).

100. Simon, H. A., and Newell, A. Models: their uses and limitations, The State of Social Sciences, N. White (ed.) (University of Chicago Press, Chicago, Ill., 1956).

101. Simons, Leo. New axiomatizations of S3 and S4, J. Symbolic Logic 18, 309-316 (1953).

102. Solomonoff, R. J. An inductive inference machine (privately circulated report) (August 14, 1956); IRE National Convention Record, Vol. 5, Part 2, 1957, pp. 56-62, Annals of Mathematical Studies No. 34 (Princeton University Press, Princeton, N. J., 1956).

103. Solomonoff, R. J. A New Method for Discovering the Grammars of Phase Structure Languages, AFOSR TN-59-110, under Contract No. AF49(638)-376, April 1959 (ASTIA AD No. 210 390).

104. Solomonoff, R. J. The Mechanization of Linguistic Learning, AFOSR-TN-246 under Contract No. AF49(638)-376, April 1959 (ASTIA AD No. 212 226).

105. Steinbuch, K. Automatic speech recognition, NTZ 11, 446-454 (1958).

106. Strachey, C. S. Logical or non-mathematical programs, Proc. Assoc. Computing Machinery (September 1952).

107. Taylor, W. K. Pattern recognition by means of automatic analogue apparatus, Proc. Brit. I. R. E. 106, Part B, pp. 198-209 (March 1959).

108. Tersoff, A. I. Electronic reader sorts mail, Electronic Industries, pp. 56-60 (July 1958).

109. Turing, A. M. Can a machine think? in J. R. Newman, The World of Mathematics, Vol. 4 (Simon and Shuster, Inc., New York, 1956).

110. Turing, A. M. On computable numbers, Proc. London Math. Soc., Series 2, 42, 230-265 (1936-37). See also a correction, Ibid, 43, 544-546 (1937).

111. Unger, S. H. A computer oriented toward spatial problems, Proc. IRE 46, 1744-1750 (October 1958).

112. Unger, S. H. Pattern detection and recognition, Proc. IRE 47, No. 10, p. 1737 (October 1959).

113. Uttley, A. M. The classification of signals in the nervous system, Memorandum 1047, Radar Research Establishment, Great Malvern, England, 1954. Also published in the EEG Clin. Neurophysiol. 6, 479 (1954).

114. Uttley, A. M. The probability of neural connections, Memorandum 1048, Radar Research Establishment, Great Malvern, England, 1954.

115. Yngve, V. H. Programming language for mechanical translation, Mechanical Translation, Vol. 5, No. 1 (July 1958).