Final Report

Covering the Period from November 1965 to May 1970

RESEARCH ON ISODATA TECHNIQUES

By: G. H. BALL D. J. HALL

Prepared for:

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
INFORMATION SYSTEMS BRANCH, CODE 437
Attention: JOEL TRIMBLE

CONTRACT Nonr 4918(00)

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

37

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

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

1. ORIGINATING ACTIVITY (Corporate author)
Stanford Research Institute
Menlo Park, California 94025

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3. REPORT TITLE
RESEARCH IN ISODATA TECHNIQUES

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

5. AUTHOR(S) (First name, middle initial, last name)
Geoffrey H. Ball and David J. Hall

6. REPORT DATE
May 3, 1971

7a. TOTAL NO. OF PAGES
36

7b. NO. OF REFS

8a. CONTRACT OR GRANT NO.
Nonr-4918(00)

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
SRI Project 5533
Final Report

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

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES
None

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research
Washington, D.C. 20360

13. ABSTRACT

ISODATA and the Singleton-Kautz algorithm, techniques for clustering multivariate data, are described and some theoretical and experimental results are given. Other techniques briefly described include: PROMENADE, an interactive graphic computer-based system for the analysis of multivariate data (PROMENADE eventually allowed us to control ISODATA interactively and to provide plots and graphs that assist in interpreting data); the Rosen-Hall discrimination algorithm, a discrimination procedure that uses a modification of the ISODATA algorithm, in combination with categories; ISODATA-Lines, a clustering technique for clustering objects around line segments rather than around points; a technique for permuting the rows and columns of a matrix in order to place either low values or high values near the main diagonal; and a program for positioning cluster centers in a two-dimensional plot in a way that approximates intercluster distance.

During the course of the project we gathered about 950 articles on clustering and wrote two survey/tutorial reports and a KWIC index to papers on clustering. We briefly explored the use of the clustering model as a structure within which to place behavioral science data.

This report contains references to all of the reports, papers, and technical notes that have been produced on this project.

Ia

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Clustering						
Pattern recognition						
Classification						
Interactive computer system						
Modelling						
Data analysis						
Statistics						
Clumping						
Numerical taxonomy						

Ik

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U S A

Final Report *January 1972*

Covering the Period from November 1965 to May 1970

RESEARCH ON ISODATA TECHNIQUES

By: G. H. BALL D. J. HALL

Prepared for:

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
INFORMATION SYSTEMS BRANCH, CODE 437
Attention: JOEL TRIMBLE

CONTRACT Nonr 4918(00)

SRI Project 5533

Approved by:

DAVID R. BROWN, *Director*
Information Science Laboratory

BONNAR COX, *Executive Director*
Information Science and Engineering Division

Copy No. 13

ACKNOWLEDGMENTS

Many people have contributed to this project:

- Professor Ingram Olkin of Stanford University--a helpful, instructive, and thoughtful consultant; a perceptive friend and counselor.

- Dr. Richard Singleton to whom we owe much of our theoretical understanding of our algorithms.

- Dr. Charles Dawson for his thorough and competent exploration of aspects of ISODATA.

- Mr. Dan Wolf and Mr. Jim Eusebio, who labored long on creating effective and efficient computer programs and who improved and added concepts and ideas throughout the project.

- Mr. Herman Friedman, Mr. Jerrold Rubin, Dr. John Sammon, and many others with whom we have discussed clustering and from whom we have learned.

CONTENTS

ILLUSTRATIONS

I INTRODUCTION

This final report summarizes work done over a five-year period on ISODATA and related topics. It relates the work initiated by this project to other research that developed out of this work. It points to other papers and reports that give the details about related methods, applications, and techniques. In this section, we give an overview of the project; subsequent sections give a more detailed description of each aspect of the project, with references to the reports and papers that deal with that aspect of the study. We conclude with our present view of the techniques developed under the sponsorship of this project, some conclusions, and some recommendations for further research. This report assumes that the reader is acquainted with clustering. The reader who is not familiar with clustering is referred to our report "Classification Analysis" for the needed background.

The project effort centered around the ISODATA technique that we developed during the Spring of 1965.* ISODATA grew out of a need for some technique that would allow us to analyze high-dimensional relationships in multivariate data. The existence of ISODATA led Dr. Richard Singleton and Dr. William Kautz to ask whether a similar technique could be developed that had a less ad hoc quality than ISODATA. This led them to develop, with the support of this project, the Singleton-Kautz algorithm for minimizing the sum of the squared distances from cluster centers. Comparison of the two techniques led us to see that ISODATA also tends to minimize the sums of the squared distances from cluster centers, although it uses a different algorithmic procedure to do so.

Analytical effort led to some theoretical results by Singleton characterizing the shape of the sum-of-squared-distance versus number-of-clusters curve.

We began systematically to search the literature for clustering techniques. This search led to the preparation of two survey/tutorial papers

* We would like to acknowledge the similarities of the ISODATA technique to that of Adaptive Sample Set construction developed earlier by Dr. George Sebestyen[1] (references are listed at the end of the report) and to that of Dr. Edward Forgy developed about the same time. The development of ISODATA was, however, an independent effort.

and of a KWIC index of about 950 papers that describe either techniques or applications of techniques. Our literature search clearly showed us that what we were doing was closely related to other work being done in different disciplines under different names.

During this time and throughout the contract, we continued to apply ISODATA and the Singleton-Kautz algorithm to real and artificially generated data. These applications indicated the need for plots, graphs, and techniques for interpreting the output from a clustering procedure and for ways to facilitate manipulation of the input data. This need led us to consider the use of an interactive graphic computer to help us control and interpret clustering of data. We obtained support from Rome Air Development Center for the PROMENADE system, which eventually allowed us easily to manipulate and transform our data, interactively to control ISODATA and the Singleton-Kautz algorithm, and to produce plots to assist in interpreting the results of the clusterings.

Several additional algorithms also resulted from our explorations. One, the Rosen-Hall Discrimination algorithm, uses category information to guide the clustering of data--basically, it clusters the data within categories, only subdividing within-class clusters when necessary to avoid confusion with data from other categories. A second algorithm, ISODATA Lines, extends the notions of ISODATA, from clustering around a point cluster center, to clustering around a line-segment cluster center. A third algorithm rearranges the rows and columns of matrices to place either larger or smaller values near the main diagonal. Such a procedure organizes the matrix, such as a distance matrix, in a way that makes it much easier to see patterns in the distribution of values, and exceptions to these patterns. A fourth algorithm was developed to position cluster centers in a two-dimensional computer plot in such a way that distance between the cluster centers is roughly preserved.

As we used and considered clustering and discrimination techniques for data analysis, we came to believe that the structures provided by these techniques could be used as a dynamic model structure for representing phenomena in the social sciences. The models would be computer-based to facilitate the user's ability to manipulate them. We have only a rudimentary understanding concerning the extent to which classification-based models can be used to develop dynamic models for social science data. The direction, so far, seems promising.

Throughout the project we have given a number of talks, written papers, and convened workshops to create the opportunity for others to become aware of the potential of clustering techniques and to try to create some conceptual framework within which to view the techniques that had been developed by a variety of people.

Finally, the ISODATA algorithm was adapted to meet the particular needs of the Navy's processing of weather data. A slight modification of the ISODATA algorithm made possible the automatic processing of satellite photographs of cloud formations for movement. From the movements of clouds over time can be estimated the wind directions and velocities over portions of the Central Pacific, where it is difficult to obtain data from surface observation. To quote from the project report:

> The basic concept is to locate a limited number of cluster centers using digitized representations of the cloud patterns in a photograph. Such centers, which are determined on the basis of area and brightness, are analogous to centers of gravity in mechanics. Measurements of displacements of these centers in successive photographs provide an indication of cloud motion. Cluster centers are found by an objective computer technique called ISODATA that was developed in an earlier program of pattern recognition carried out at SRI. Cloud motions are derived by a separate program that matches centers found on two pictures. In a number of cases studied, the computed cloud motions were reasonable and agreed qualitatively with motions perceived visually from time-lapse presentations of the same data.[2]

We will now expand somewhat the preceding summary regarding the most important aspects of the work that we have performed and indicate how the reader can obtain more detailed information should he wish to do so. We will take each of the aspects of the project in turn and more or less in the order in which it was developed during the project. (The exception is the Rosen-Hall Algorithm. We will not discuss it further, since a written description is readily available.[3])

II ISODATA

ISODATA, Interactive Self-Organizing Data Analysis Technique A,[4,5,6,7] is a simple, straightforward ad hoc procedure to sort multivariate data in clusters such that the objects within each cluster are more like the average object--the cluster center--in that cluster than they are like the cluster center in any other cluster. The process successively reduces the sum of squared distances (SSD) around the cluster centers by iteratively repeating two steps: (1) assigning all objects to that cluster center to which they are nearest (as measured by some measure of distance, usually Euclidean), and (2) shifting the cluster center to the average of all of the multivariate objects that have been assigned to that cluster by Step 1. Both steps reduce the sum of squared distances around closest cluster centers. Since neither step in the process can cause the sum of squared distances to increase, we can say that the convergence process is direct. The process will reach a stable partition in a finite (usually small) number of iterations for a finite data set (since each step, until the final step, reduces squared error by some positive amount and since squared error is bounded from below by zero). Unfortunately, the resulting partition depends on the initial positions of the cluster centers, as we discuss below.

A. Splitting and Lumping

We added to ISODATA lumping and splitting procedures that decreased or increased the number of clusters.

Lumping combines clusters that are sufficiently close, relative to a threshold set by the user. The average of the objects in the two clusters becomes the new cluster center. Splitting divides a single cluster into two clusters. It involves first the evaluation of the desirability of dividing the cluster into two clusters and second, a procedure for doing this splitting.

In the original ISODATA algorithm, splitting was performed by evaluating each cluster on the basis of whether the maximum standard deviation along any one of the dimensions for each of the clusters exceeded a control parameter, θ_E. If θ_E was exceeded, the cluster was split. Certain problems result from using this procedure. In particular, it is possible to select the value of θ_E such that a cluster is split and then at a later time have the two resulting clusters recombine because the distance between the means of the two resulting clusters as measured over all variables was too small relative to the value of θ_C, the parameter that controls when two clusters are to be recombined.

The procedure now programmed with the ISODATA algorithm performs a trial splitting for each of the clusters. This splitting criterion functions as follows:

(1) Find that one variable among the variables of the data having the largest standard deviation about the mean of the cluster.

(2) Sort the data into two subsets--one subset consisting of all patterns having a value larger than the average value over all objects in the clusters in that one variable and another subset consisting of all patterns having values smaller than the average in that one variable. (Note that a comparison of the value of one variable with the threshold is all that is required for this step, which requires little computation.)

(3) Find the averages of these two subsets.

(4) Use the distance between these two averages as an approximation to the distance that would exist between the two cluster centers resulting from the split after one more iteration. It is an approximation because the effect of patterns in other clusters is not taken into account.

(5) Compare this magnitude with the threshold (1.1) θ_C and split the cluster if that threshold is exceeded. The threshold θ_C is the parameter that determines when two clusters are to be combined into a single cluster (lumping). The advantage of the new splitting criterion is that it uses the distance between the new cluster means after splitting using all of the variables rather than just evaluating the cluster on the basis of the largest standard deviation in any one variable. It also makes possible, although this has not been implemented, the selection of that cluster that will maximally decrease the sum of the squared distance when split. This is useful if the ISODATA algorithm is used to trace out the curve of minimum squared distance (MSD) versus the number of clusters, as is done in the Singleton-Kautz algorithm.

The recombining or lumping of two clusters depends on measuring the Euclidean distance between all pairs of cluster averages and comparing this distance with a threshold, θ_C. Currently, all clusters having interpair distances greater than θ_C have been recombined. In the future it might be desirable to combine that single pair of clusters that minimally increases the squared distance. This would be simple to do because the sum-squared distance is a function only of the distance between the two cluster centers that are being considered for recombination, and the number of patterns in each cluster.* If this were done, it would result in the complete elimination of the process parameters that control the ISODATA process.

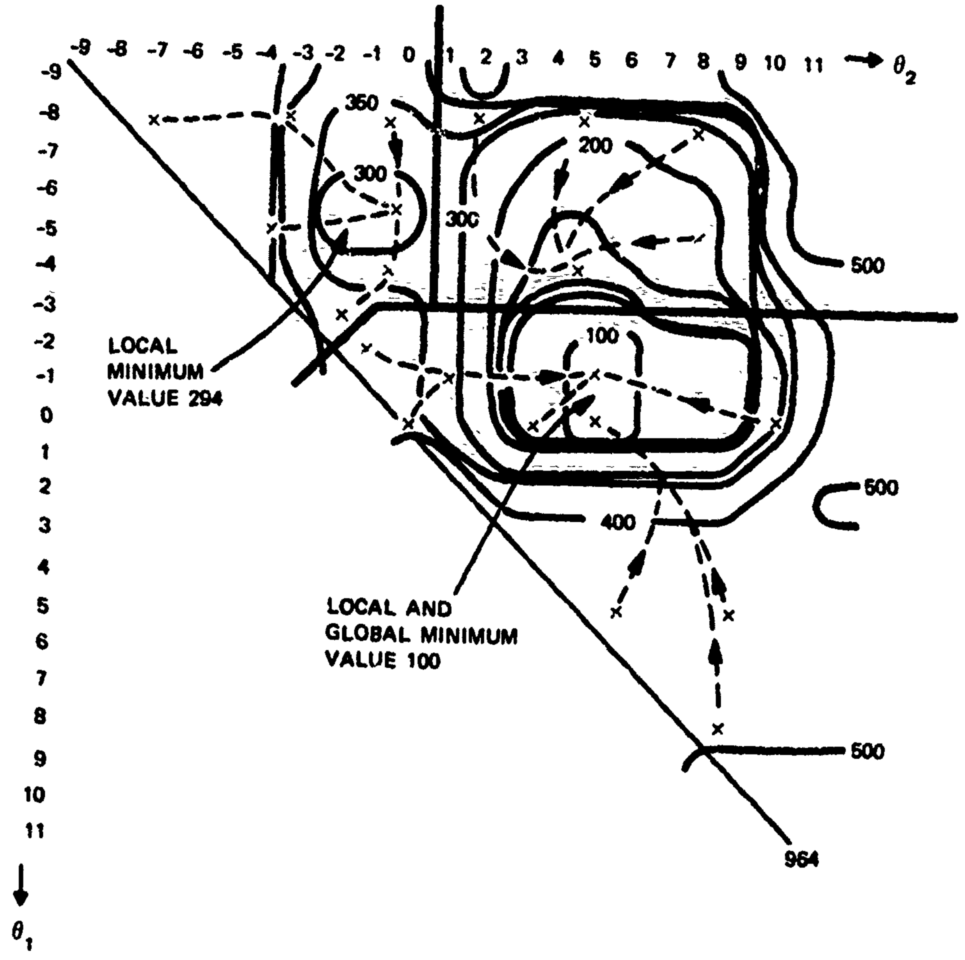
In certain cases it seems that removal of these parameters from consideration would be useful. In other situations, when we wish only to use the magnitude of the distance between the cluster centers to determine the number of clusters, it may be desirable to retain θ_C.

B. Convergence Properties of ISODATA

From experimental evidence we know that ISODATA does not necessarily find the overall minimum of the sum--it will, however, stabilize either at a local minimum of the sum or at a sufficiently extensive flat region of the criterion function. The representation of a one-dimensional data set as a contour map of SSD versus the values of the partitioning thresholds, θ_1 and θ_2, allows us to investigate the dynamics of the ISODATA process (without splitting or lumping). This plot is shown in Figure 1. This representation gives the value of the sum of the squared distance as a function of the position of two thresholds placed along the real line for the data shown in Figure 2. In using this representation we use the knowledge that the convex hulls of an MSD partition cannot intersect. The tracks shown on the contour plot show how the ISODATA algorithm shifted thresholds from iteration to iteration. We see that this "settling process" does not always find even a local minimum of the sum-squared-distance surface but that it may stop on a "shelf" in the SSD function fairly remote from a local minimum point of the sum-squared-distance surface.

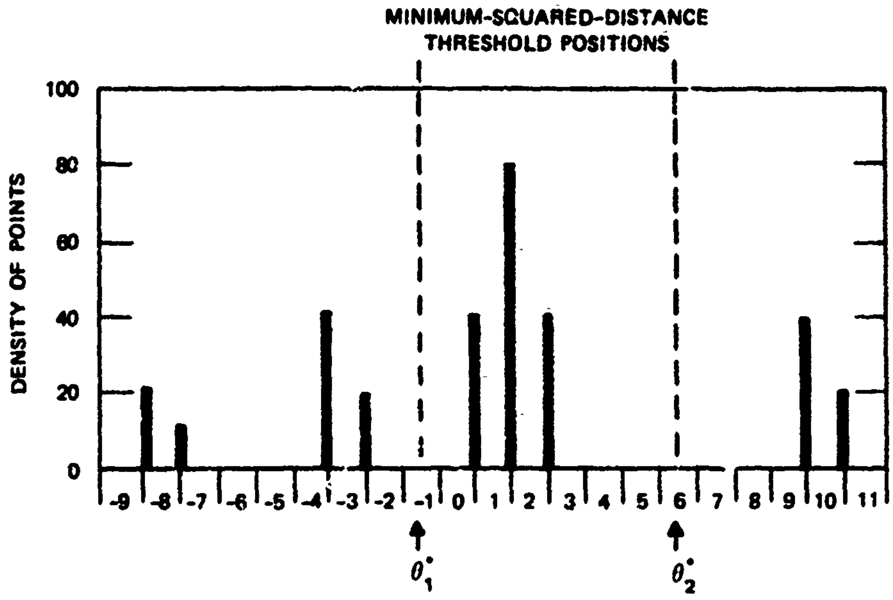
The contour plot also illustrates the existence for this data set of two minima of the sum-squared-distance function for three clusters. Using the plot, we have obtained examples of one or two minima for two clusters and one or two minima for three clusters.

* This is strictly true only until the next iteration.



TA-658582-2

FIGURE 1 CONTOUR PLOT OF VALUES OF SSD FOR TWO THRESHOLDS, θ_1 AND θ_2



TA-658582-4

FIGURE 2 ONE-DIMENSIONAL PATTERNS

C. Measures of Similarity Used with ISODATA

The ISODATA partitioning can be and has been done with respect to a variety of measures of similarities of objects to cluster averages. The measures of similarity used thus far are:

(1) Normalized dot products between objects $\{x\}$ and cluster averages $\{m\}$, where the normalization is with respect to the magnitudes of the means and the data points. This can be expressed as $(\vec{x} \cdot \vec{m})/||(\vec{x})|| \cdot ||(\vec{m})|| = \cos\left(\underline{\vec{x}, \vec{m}}\right)$.

(2) The dot product between the object and the cluster averages. This can be written as $\vec{x} \cdot \vec{m} = ||\vec{x}|| \cdot ||\vec{m}|| \cos\left(\underline{\angle \vec{x}, \vec{m}}\right)$.

(3) Euclidean distance squared. This can be written as $||x - m||^2 = x \cdot x - 2x \cdot m + m \cdot m = (x - m) \cdot (x - m)'$.

(4) Mahalanobis distance, which includes Euclidean distance as a special case, which can be written as $(x - m)\, W^{-1}\, (x - m)'$, where W is usually the pooled covariance matrix, or the sum-of-products-within matrix.

Originally ISODATA was programmed for binary data and for dot-product measures. Then we programmed ISODATA to use Eculidean distance. Under the stimulus of the work by Friedman and Rubin,[8] we developed a version of ISODATA that used Mahalanobis distance with the covariance matrix used being either the pooled covariance matrix around cluster centers, or the diagonal covariance matrix computed for each cluster independently.* We found that for our purposes an initial normalization of the standard deviations of each variable was sufficient and for this reason did not pursue the use of the more elaborate measure of distance.

* Using the off-diagonal values of the covariance matrix derived from a single cluster would tend to cause instability unless each cluster had a substantial number of objects assigned to it.

III THE SINGLETON-KAUTZ ALGORITHM

The Singleton-Kautz algorithm[2] was developed by Dr. Richard K. Singleton and Dr. William H. Kautz of Stanford Research Institute in 1965. This algorithm seeks explicitly to minimize the sum of the squared distances around cluster centers. The algorithm uses the following steps to perform this minimization:

(1) All objects are assigned to a single cluster.

(2) The object farthest from the mean of the single cluster is assigned to a second cluster.

(3) Each object is tested to determine whether its reassignment to the second cluster will reduce the sum-squared distance (SSD). (Fortunately the computation requires only the evaluation of the <u>change</u> resulting from the reassignment.)

(4) When it is no longer possible to reduce the SSD by reassigning any single objecτ to a different cluster, then the number of clusters is increased by one, and the process is repeated. For data sets of 200 objects, experience indicates that about four cycles through the set of objects lead to the situation in which no single object's reassignment will result in a reduction of the SSD. For larger numbers of objects, the number of cycles may ncrease considerably. The amount of cycling required also depends on the "lumpiness" of the structure of the data.

(5) The number of clusters is <u>increased</u> to some maximum number of clusters. This maximum number is set by the person using the program.

(6) When the limit is reached, then the cycle is reversed and the number of clusters is changed by <u>reducing</u> the number of clusters by combining those two clusters that <u>minimally</u> increase the SSD. After combining those two clusters, the cycle described above is then used to attempt to further decrease the SSD. If the SSD found is smaller than the SSD found in any previous cycle, then the partition obtained on this new

clustering is substituted for the partitioning found the previous time.

(7) This increasing and decreasing of the number of clusters is continued until it is not possible to reduce the SSD further, or until a user-specified number of iterations have been performed. At this point the process terminates.

Critical steps in this process are the selection of the data point used to initiate a new cluster, the ordering of data points, and the choice of those two clusters that are to become combined when the number of clusters is reduced. These comments can be summarized by saying that the choice of the starting points for the iterative hill-climbing to a MSD partition determines whether the partition obtained is the minimum among all local minimum-squared-distance partitions of the data set.

IV COMPARISON OF ISODATA AND SINGLETON-KAUTZ

The Singleton-Kautz algorithm and ISODATA both produce a disjoint partition of the data set with relatively similar objects being placed in the same partition. Similarity is measured by distance to a cluster average. Variations of these techniques can be obtained in the case of the Singleton-Kautz algorithm by modifying the criterion against which improvement in the partitioning is measured, and in the ISODATA technique by modifying the measure of similarity and by modifying the procedure by which clusters are split and lumped. Global or local evaluating criteria can be used with ISODATA to further constrain the solution obtained. No explicit distributional assumptions are made in either of these techniques. However, it is assumed that the distance measure or the criterion used is adequate to reflect the structure of the data.

These cluster-seeking techniques are most effective in describing situations in which clusters of data exist. These techniques are not particularly efficient in describing data having relatively uniform random variability within a low-order linear subspace of the original data space. For such situations factor-analytic techniques that look at these linear subspaces seem more appropriate. The clustering techniques can, however, still be used in these situations to provide empirical data categories. The cluster-seeking techniques try to group patterns so that the average squared distance from cluster averages is minimized. Factor-analytic techniques find the factors--composite variables--that define a lower-dimensional space and then retain the full variability of the data within that lower-dimensional space.

Both the Singleton-Kautz algorithm and ISODATA are invariant with respect to orthogonal transformations and translations of the data. Both cluster-seeking techniques, when using either a criterion or a measure of similarity corresponding to Euclidean distance, are sensitive to changes in scaling, although they are not sensitive to rotations of the data or the position of the origin. Changes in the data set that affect normalizations based on the data sets, such as the standard deviation about an average, may modify the clustering obtained. When the ISODATA technique is used with the Mahalanobis distance, it is relatively insensitive to the scaling of the data.

Invariance is desirable if it removes irrelevant variation. If the particular variability is important, then a technique has to be developed that is sensitive to this variation. For example, if scale is important and there is some natural way of defining the scale, or if there is a desire to weight certain variables more heavily, then invariance with respect to scale would not be a desirable feature for a technique.

ISODATA appears to be directly extendable to the clustering of points around line segments or planar sections. The ISODATA-Lines algorithm clusters points around line segments. This algorithm uses the following notions from ISODATA:

(1) The creation of new cluster centers (the cluster centers are now line segments).

(2) The evaluation of the usefulness of a given line segmen..

(3) The iterative shifting of the line segment to place it in "better" position where the new position is primarily dependent on objects close to the line segment (i.e., not averaging over all objects, but rather over a subset).

(4) The combination of those line segments that can be combined without greatly reducing the information about the structure of the data.

We investigated the desirability and the feasibility of clustering data around triangular planar sections. While this would enable us to approximate mixtures of nonlinear, two-parameter surfaces that are embedded in a hyperspace, the amount of data needed to obtain somewhat stable positioning of the triangular planar sections appears excessive.

The computation times for the "inner loop" of the Singleton-Kautz algorithm and for the "inner loop" of the ISODATA program using Euclidean distance as its measure of similarity are approximately equal. If a more complicated distance function is used, such as one of the Mahalanobis type distances, and _if_ the Singleton-Kautz algorithm inverts a matrix after each data sample, the computation time of its "inner loop" would be greater than that of an ISODATA procedure that only recalculated the matrix after each iteration.

Experimentally on 225 two-dimensional data points we observed that the Singleton-Kautz algorithm finds a partition of the data that has an MSD that is about 10 percent lower than that of the partition found by

ISODATA.* We do have instances, however, when ISODATA has found a lower MSD.

The partitions were almost the same in most instances. The Singleton-Kautz algorithm is considerably easier to run, since it systematically provides values for minimum SSD for all numbers of clusters up to a value KMAX set by the user. However, in runs on higher dimension data that had a considerable number of wildshots, ISODATA proved to be easier to interpret and run since it was not as affected by the wildshots. That is, ISODATA rapidly increased the number of clusters until the wildshots were isolated. In this latter application, ISODATA was more effective.

For systematic analysis of relatively clean data, where finding the MSD partition for small numbers of clusters is a reasonable goal, the Singleton-Kautz algorithm appears to find partitions that have lower values of SSD than ISODATA. From past experience with other data, ISODATA appears to be superior for noisy data, where the goal is quick isolation of the principal modes of the data with exclusion of outliers.

The program implementing the Singleton-Kautz algorithm is easier to use in a batch-processing computer. We feel that ISODATA proved more versatile and as easy to use in an interactive computer in which the judgment of the operator is used in lumping, splitting, and evaluating clusters.

The speed of convergence of the two algorithms to an MSD partition apparently depends on the number of patterns and number of dimensions.

* The lowest value of SSD may not produce the partition of the data that is easiest to interpret. For example, the minimization may give too much weight to outliers--objects distant from any cluster.

V THE SHAPE OF THE MSD CURVE

Dr. Richard Singleton has been able to show[8] by counterexample that the curve displaying the MSD versus the number (K) of clusters is not convex. He has been able to show, however, that while the curve is not convex with respect to all possible pairs of points, it exhibits convexity with respect to those pairs of points having as one member of the pair either K = 1 or K = N, where N is the number of data points. This form of weak convexity has been described previously and labeled "star-shaped."[9,10] It is worth noting that, at least in appearance, the weakening of the convexity of this curve to star-shaped form does not appear to allow the MSD-versus-K curve to be very nonconvex.

The star-shapedness of the MSD-versus-K curve can be used in evaluating an empirically obtained MSD-versus-K curve by testing the empirical curve for star-shapedness. When, for a particular value of K, the MSD (K) violates this star-shaped condition, we would attempt to find a new partition such that the curve becomes star-shaped.

Dr. Richard Singleton has shown that for a minimum sum-of-squared-distance (MSD) partition it is necessary (but not sufficient) for the hyperplane that is the perpendicular bisector of the line connecting any two cluster averages not to intersect the convex hulls of those two clusters. (This requires that the convex hull* of the objects in one cluster not intersect the convex hull of objects in any other cluster.)

It follows from this condition that a partition can be a stable MSD partition only if the averages of the respective clusters are such that the above condition is satisfied. The ISODATA procedure uses this condition to seek an MSD partition by reassigning objects that do not meet that condition to that cluster having the closest cluster center. The use of the perpendicular bisector can be generalized to distances measured using Mahalanobis-type distance.

* The convex hull is the minimum volume convex body sufficient to contain all of the objects in one cluster.

VI SURVEY AND TUTORIAL EFFORTS

During the course of this project we wrote two survey/tutorial papers: one[11] on clustering in 1965 and one[12] on clustering and discrimination techniques (with emphasis on clustering) in 1970, as well as a number of tutorial papers related to these surveys. During this time about 950 references were collected and placed on cards and annotated.[13] A KWIC program was modified to allow sorting on a set of descriptors.

The earlier survey will be of most use to the reader for its description of a number of algorithms. The later survey attempts to give an integrative model of clustering and discrimination techniques. It includes a number of comments related to interpretation and to assumptions and questions concerning the use of clustering. Rather than attempt to summarize this report, we would suggest that the reader obtain a copy of these surveys for his own use.

VII APPLICATIONS OF ISODATA AND SINGLETON-KAUTZ

Both algorithms have been distributed to various users around the country. In general, users have found that on batch processing computers, Singleton-Kautz is the easier of the two to use and it has been used extensively by a number of people. (Copies of the Singleton-Kautz program and of program notes are available from the authors.) As a result of application of these techniques, it became very clear that we needed output from the programs that would help us interpret the results. (Both programs have more words of code devoted to organizing and printing out the results of the clustering than they do words of code concerned with the actual clustering.) The output from ISODATA concentrates on describing the cluster centers and the relationships between the cluster centers. The Singleton-Kautz algorithm uses notions from the analysis of variance to give indications of the importance of various variables and to assist the user in plotting the sum-of-squared-distance curve versus the number of clusters. The Singleton-Kautz printout also provides the data needed to plot the same curve for individual variables. ISODATA provides a distance matrix giving the distances between the cluster centers. This matrix is helpful in relating the specific clusters to each other.

Out of our need to interpret the results from clusterings and the availability of an interactive graphic computer came the development of PROMENADE--an interactive graphic multivariate data analysis system.

VIII PROMENADE

PROMENADE is described in considerable detail in its project report,[14] and from the viewpoint of the user in a series of papers.[15-19] Most of the PROMENADE work was supported by Rome Air Development Center. The particular work on PROMENADE supported by this project dealt with giving the user of the system an ability to control interactively the ISODATA algorithm, to select and transform his data, and to use a variety of plots to help him interpret the results he obtained from clustering.

Interactive clustering is described in a paper[19] written in 1969. To give the user interactive control, we needed to provide him with information needed to control the process intelligently. We chose to give him this information in three ways:

- By viewing the parameter values on a display.
- By viewing the link-node graph that shows cluster centers and links between the cluster centers.
- By viewing the profile or waveform plot that shows an object as a waveform of connected line segments with variable values plotted on a set of parallel axes.

The object of interactive clustering is to allow the user to control the detailed workings of the algorithm so that he can tailor the algorithm to his data. The kinds of things that a user might want to control are: the number of objects in a cluster (perhaps as some function of the dispersion of the cluster); the total number of clusters; which clusters are fixed and which are allowed to change during a particular iteration; which objects are used within a particular iteration; and which variables are used in calculating distance. He needs help from the system in keeping track of what he has done. He needs an indication, which we believe is given most effectively by graphs and plots, of the effect that his actions have had.

Before the end of the project we were able to program some of these operations, such as the following: use the data manipulation section of PROMENADE to store data sets that we used or created, and select out the variables and objects that we wished to use; delete selected objects during the clustering either by sequence number or by association with a given cluster; transform variables or produce composite variables if we desired; and control the clustering while watching either the link-node or the waveform plots.

The user primarily has control over whether to lump, to split, or to do neither; he can also specify which clusters are to be split or lumped together.

We look forward to having an opportunity to develop much more complete control over the clustering. For example, we would like the user to be able from waveform or link-node plots to:

(1) Create, delete, freeze (set aside), or use cluster centers

(2) Simply iterate to converge to a stable partition

(3) Lump or split specified clusters

(4) Delete, freeze, or use specified objects

(5) Ignore or use specified variables

(6) Use the electronic pointer on the display scope to specify which cluster(s) or variable(s) he wants to act on.

(7) Plot only selected objects--perhaps only those objects in one or two clusters

(8) Show before and after positions of cluster centers--at least the positions as projected onto a two-dimensional space.

For a variety of reasons, mostly lack of funds, we were not able to experiment with the controls over the clustering program. Based on our limited experience, however, we would recommend further investigation of interactive clustering.[20]

IX ISODATA-LINES

ISODATA-Lines, a computer program developed by James Eusebio and Geoffrey Ball,[21] attempts to find, in a set of multivariate data points, subsets that may be represented by piecewise-linear curves. In effect, the program partitions the data into clusters around line segments or "cluster axes." Such axes may be simple in structure or they may be complicated networks of lines with many branches. This program is an extension of the ISODATA program, which also forms clusters of objects but which represents each cluster by a single point.

Both programs allow a number of objects to be described by a smaller number of cluster centers. It is frequently easier to gain an understanding of the structure of the data by examining just the cluster centers than by examining the full data set. Examples of this are shown in the projection plots described below. These techniques are esp.cially useful for showing the structure of high-dimensional data, where a visual picture of the objects is not available.

In practice, ISODATA-Lines first partitions the data into clusters, each of which is represented by a cluster center. Then certain pairs of cluster centers are linked by straight lines to form piecewise-linear curves. (We call linked cluster centers "cluster nodes," to distinguish them from "cluster links.") These links divide the set of cluster nodes into disjoint chains. ISODATA-Lines can be used to discover "curvilinear structures" in a data set.

ISODATA-Lines can aid the visual interpretation of high-dimensional data. If the given data set has more than three dimensions, no direct visual picture of it is available. Even two-dimensional pictures of three-dimensional data can hide much of the data's structure. One procedure for picturing high-dimensional data is as follows: first, perform a cluster analysis to reduce the number of objects to be plotted; then project the cluster centers onto some plane. But this is not enough, because distance relations between cluster centers are generally distorted in projecting onto a plane. Cluster centers whose projections on a plane lie close together may, in fact, be far apart in the original high-dimensional space. ISODATA-Lines links pairs of cluster centers which are close in the high-dimensional space, and between which there are objects. A projection plot that includes the projections of these links indicates clearly which cluster nodes are near one another.

For example, Figure 3 shows the projection onto a plane of a set of three-dimensional data. This artificially generated data set consists of objects scattered around two parallel helical arcs plus a small "blob" of objects to one side of the helices. This data set was processed by ISODATA-Lines, and the resulting cluster nodes and cluster links were plotted as shown in Figure 4. The two separate arcs show up clearly, though from this projection alone we cannot conclude that they are helical.*

Possible future improvements to the algorithm include: (1) an improved test to decide whether two cluster centers should be linked, (2) lumping and splitting of clusters, as is done in ISODATA, and (3) new ways of partitioning the data set which depend on the distances of points from links rather than on the distances of points from nodes.

* A third dimension can be added to these projection plots by marking the nodes with symbols (e.g., squares) which vary in size with the distance of the nodes from the projection plane.

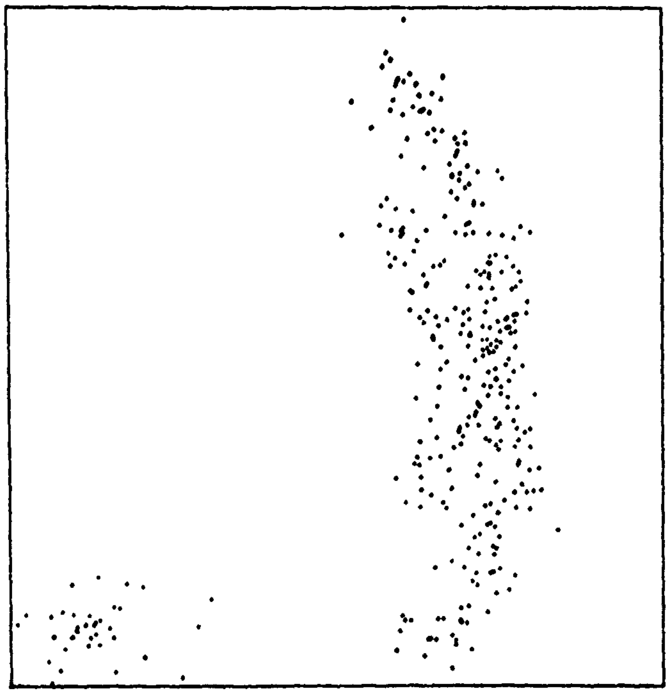


FIGURE 3

PROJECTION ONTO A PLANE OF A SET OF THREE-DIMENSIONAL HELICAL DATA

TA-5533-51

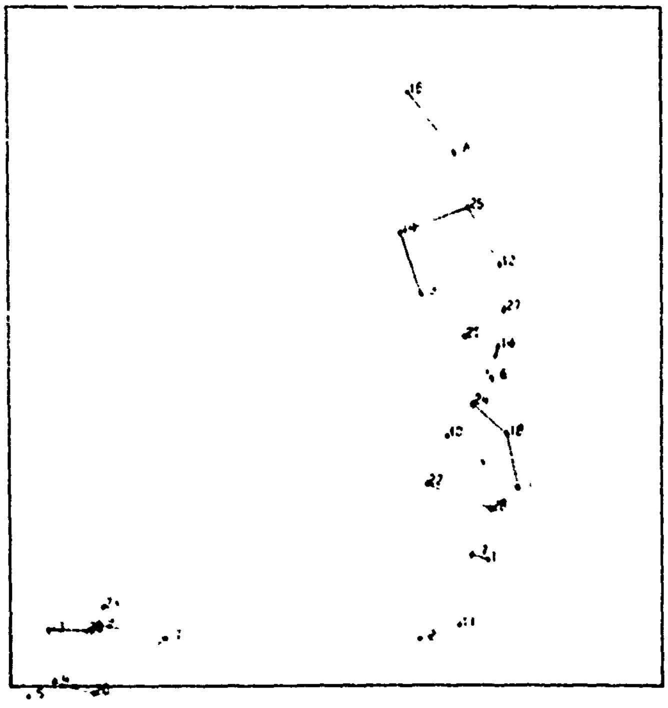


FIGURE 4

PLOT OF CLUSTER NODES AND CLUSTER LINKS RESULTING FROM PROCESSING THE DATA SET OF FIGURE 3 BY ISODATA-LINES

TA-5533-52

X COMPUTER OUTPUT MANIPULATION

The analysis of multivariate data using the digital computer can produce vast quantities of output. The arrangement of the output that the human analyst will be asked to interpret is extremely important. For example, an analyst looking at a covariance or correlation matrix will frequently first attempt to find the large values in this matrix. His task can be made markedly easier by algorithms that automatically place all large values near the main diagonal. This convenience can be dramatically helpful in interpreting the results of processing the data.*

The matrix permutation program permutes the rows and columns of a matrix of pairwise distances between vectors to maximize† the criterion

$$\text{CRIT} = \sum_{i=1}^{K} \sum_{j=1}^{K} m_{ij}^{2} (i - j)^{2} \quad ,$$

where m_{ij} is the ij^{th} element in the permuted square matrix and K is the number of rows (and also the number of columns). This is maximized over values obtained by sequentially permuting rows and the corresponding columns.

A second program produces a line printer plot of a set of cluster centers produced by a cluster-seeking program. The cluster centers are considered nodes in an undirected graph. This program positions these nodes in a way that tries to put close (in the sense of Euclidean distance) nodes near each other on the page. The virtue of the program is that it is simple and fast. The program can be used to provide the user

* See Ref. 22 for a related technique for binary matrices and for references to other papers.

† The program can easily be modified so that the sum is minimized, which results in large values being placed near the main diagonal. If the matrix given to the second program is a covariance matrix, then the rearrangement of the rows and columns to place large values near the main diagonal will place related variables close to each other.

of a cluster-seeking program some idea of the relative relationships between the clusters. It can also be used to provide an initial position for a more elaborate positioning algorithm such as Shepard-Kruskal's multidimensional scaling algorithm. It can also provide initial positioning for interactive manipulation of the graph by a user on a system such as PROMENADE.

Extensions of the notions expressed above of organizing information to other kinds of output should be equally helpful. For example, the rearrangement of variables in the output of a factor analysis could bring together all variables weighted heavily by a particular factor with the variables ordered by their weights. It seems that this human factor aspect of computer science pertaining to information arrangement and organization should be examined more closely than it has been in the past.

XI CLASSIFICATION AND COMPUTER MODELS FOR THE SOCIAL SCIENCES

In working on ISODATA and PROMENADE, we came to feel that there was much fertile territory for developing social science models based on clustering and on discrimination techniques. Classification (that is, clustering and discrimination) provides a suggestive vantage point from which to view some social science phenomena, for it suggests some alternatives to models that have been used to date.

The concepts and descriptive structure provided by clustering, combined with our experience on interactive graphic computers, cause us to think that it would be possible to provide social scientists with an interactive computer-based system for storing and examining dynamic processes. This system would complement classification systems such as PROMENADE and would provide a means of capturing details obtained through detailed data analysis. One major aim of such a system would be to allow the user easily to retrieve and modify pertinent information, while allowing the observer operator to use his intuition. Criticism could be directed to details in the model, rather than to some general overview. The consequences of changes in the model could be quickly seen.

XII CONCLUSIONS

From our present perspective, we see ISODATA as a simple, straightforward clustering technique that is somewhat awkward to use in a batch-processing mode but nearly ideal for interactive clustering. The Singleton-Kautz program provides a more explicit algorithm optimizing a well-understood function and is extremely well implemented in the existing program, which has been used and tested by a number of people. ISODATA-Lines is interesting and perhaps useful in specific applications--particularly in developing pictorial presentations of high-dimensional data having curvilinear structure (e.g., speech). Interactive clustering would appear to be most worth developing further--particularly for exploratory analysis performed in combination with use of a number of other techniques.

In a broader perspective, it appears that there is a need for more effective computer languages that allow for construction of clustering procedures. Succinct algorithmic languages are needed in order to provide a common language for communicating the nature of various clustering algorithms among users.

Finally, we believe that the reader should keep in mind that clustering techniques do not somehow produce objective Truth. Clustering techniques are explicit but, depending on a variety of factors, may produce different partitions on the same data. This explicitness is, however, of value since it makes it easier to see the assumptions and the weightings that a researcher employs in reaching his conclusions.

REFERENCES

1. G. S. Sebestyen, Decision Making Processes in Pattern Recognition (MacMillan Company, New York, New York, 1962).

2. R. M. Endlich, D. E. Wolf, D. J. Hall, and A. E. Brain, "A Pattern Recognition Technique for Determining Cloud Motions from Sequences of Satellite Photographs," Final Report, Contract N62306-69-C-0312 SRI Project 7989, Stanford Research Institute, Menlo Park, California (April 1970).

3. C. A. Rosen and D. J. Hall, "A Near Optimum Categorization Algorithm for Certain Artificial ͺaussian Data," IEEE Trons. on Information Theory, Vol. IT-12, No. 2, p. 277 (April 1966).

4. G. H. Ball and D. J. Hall, "ISODATA, A Novel Method of Data Analysis and Pattern Classification," Stanford Research Institute, Menlo Park, California (1964).

5. G. H. Ball and D. J. Hall, "ISODATA, An Iterative Method of Multivariate Data Analysis and Pattern Classification," IEEE International Communications Conference, pp. 116-117 (1966).

6. G. H. Ball, and D. J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioral Science, Vol. 12, No. 2, pp. 153-155 (1967).

7. D. J. Hall and G. H. Ball, "ISODATA, A Clustering Technique," Technical Report, Stanford Research Institute, Menlo Park, California (May 1971).

8. H. P. Friedman and J. Rubin, "On Some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, Vol. 62, No. 320, pp. 1159-1178 (1967).

9. R. K. Singleton, "A Clustering Technique for Minimizing the Sum-of-Squared Errors," Technical Note, Stanford Research Institute, Menlo Park, California (May 1971).

10. A. M. Bruckner and E. Ostrow, "Some Function Classes Related to the Class of Convex Functions," Pacific Journal of Mathematics, Vol. 12, No. 4 (1962).

11. G. H. Ball, "Data Analysis in the Social Sciences," Proceedings of the 1965 Fall Joint Computer Conference, pp. 533-560 (1965).

12. G. H. Ball, "Classification Analysis," Technical Report, Contract Nonr 4918(00), SRI Project 5533, Stanford Research Institute, Menlo Park, California (November 1970). To be published in Statistics, ed. Michael Haas (Northwestern University Press, 1971).

13. G. H. Ball and S. M. Peterson, "A Description-based KWIC of Articles on Clustering," Technical Note, Stanford Research Institute, Menlo Park, California (January 1972).

14. D. J. Hall et al., "PROMENADE--An Improved Interactive-Graphics Man/Machine System for Pattern Recognition," Final Technical Report No. RADC-TR-68-572, Contract No. F30602-67-C-0351, SRI Project 6737, Stanford Research Institute, Menlo Park, California (June 1969), AD692752.

15. D. J. Hall et al., "PROMENADE--An Interactive Graphics Pattern-Recognition System," Proceedings of the IFIP Congress 68, pp. 951-956 (August 1968). Also IFIP Congress 68, Final Supplement, Booklet J, pp. J46-J50 (August 1968).

16. D. J. Hall, G. H. Ball, and D. E. Wolf, "PROMENADE--An Interactive Graphic Computer System for Sorting Multivariate Data into Groups,' Proceedings of the Purdue Centennial Year Symposium on Information Processing, Vol. II, pp. 423-445 (April 1969).

17. G. H. Ball, D. J. Hall, and D. E. Wolf, "PROMENADE--A CRT/Console-Controlled Computer System for Analyzing Multivariate Data," in Pattern Recognition, ed. L. N. Kanal (Thompson Book Company, Washington, D.C., 1968).

18 D. J. Hall, G. H. Ball, D. E. Wolf, and J. W. Eusebio, "PROMENADE--A System for On-Line Pattern Recognition," The Future of Statistics, ed. Donald G. Watts, pp. 390-413 (Academic Press, New York, New York, 1968).

19. D. J. Hall, G. H. Ball, and D. E Wolf, "Interactive Graphic Clustering Using the PROMENADE System," Proceedings of the 1969 Social Statistics Section, American Statistical Association, pp. 65-73 (1969).

20. G. H. Ball and D. J. Hall, "Some Implications of Interactive Graphic Computer Systems for Data Analysis and Statistics," Technometrics, Vol. 12, No. 1, pp. 17-31 (February 1970).

21. J. W. Eusebio and G. H. Ball, "ISODATA-Lines--A Program for Describing Multivariate Data by Piecewise-Linear Curves," Proceedings of the International Conference on Systems Science and Cybernetics, University of Hawaii, Honolulu, Hawaii, pp. 560-563 (1968).

22. Seymour Spilerman, "Structural Analysis and the Generation of Sociograms," Behavioral Science, pp. 312-318 (July 1966).

TALKS GIVEN

In addition to those talks listed in the References, the following talks were given:

1. "Some Strengths and Weaknesses of the ISODATA Cluster-Seeking Techniques in Generating Typologies," given at meetings of the American Psychological Assoc., 2 September 1968. (Talk only.)

2. "On the Status of Applications of Clustering Techniques to Behavioral Sciences Data," (with H. P. Friedman), *Proc. Social Science Section*, Amer. Stat. Assn. Meetings, 20 August 1968.

3. "Sorting Things into Groups--New Techniques for Old Problems," Given at the American Psychological Assoc. Meetings, 1 September 1968.

4. "Sorting Things into Groups," Talk given at the Gordon Conference on Statistics in Chemistry, July 7-11, 1969.

5. "Clustering Techniques--An Introductory Statement," Geoffrey H. Ball, *Proceedings of the American Society for Quality Control*, 23rd Technical Conference, May 5-7, 1969.