ON THE STATUS OF APPLICATIONS OF CLUSTERING TECHNIQUES TO BEHAVIORAL SCIENCES DATA

Geoffrey H. Ball, Stanford Research Institute

Herman P. Friedman, International Business Machines Corporation

1.0 Introduction

What is clustering? One definition is provided by Professor David Wallace,

> "Clustering methods form a loosely organized body of techniques for the analysis of [multivariate] data. As with most methods of data analysis the aim is to find, describe, and hopefully to lead to an explanation of simple structure in a complex mass of data. Clustering methods are distinguished by the type of structure that is sought." D. Wallace (1968)

What kind of structure is sought by clustering techniques? It might be data having a structure that can be usefully described as a mixture of normal distributions with covariance matrices that are:

(a) all the identity matrix

(b) all equal but not equal to the identity matrix

(c) different for each of the underlying distributions.

For unequal covariance matrices, the data might take on the appearance shown in Figures 1 and 2. (Such highly artificial data sets are useful because they provide convenient methods of testing the characteristics of an existing or proposed clustering technique prior to its use on data having unknown characteristics). More generally distributions are not normal, as for example, in the data describing luminosity as a function of star temperature shown in Figure 3.

In the two and three dimensional data that we have shown this far, it is possible to plot the data, to see its structure and from this visual observation to reach conclusions regarding its cluster structure. (However, mixtures of even straightforward multivariate normal distributions do no easily reveal their structure when the variances are so large that the distributions overlap). In p dimensions it is not possible to see the p-dimensional relationships, and it is for this reason that the data analyst must rely on techniques such as clustering to find the structure implied by the numeric values of the data samples.

There are a large body of "clustering" procedures that take as their basic data a matrix of distances or similarities between each pair of objects. A basic problem here is the appropriate definition of this measure of similarity. This is particularly significant since the choice of distance implicitly defines one aspect of the structure. Here the work of multidimensional scaling as exemplified by the works of Shepard (1962), Kruskal (1968), Torgerson (1968) and Guttman (1968) are relevant. Their work enables one to take non-metric data for which similarities or distances have been specified and produces a metric space of quantitative dimensions in which the clustering procedures may search for group structure. The interplay between these two techniques is beginning to be recognized and is discussed by Torgerson (1968) and Green et al (1968).

Actually we believe there is an inherent circularity in talking about finding structure in data that is brought about by a looseness in language. It appears to us that the questions to be answered are akin to those in using models in science. We no longer ask for the correct (absolute) model. We do ask when is a model better than another with respect to powers of prediction, explanation, and simplicity. In the same way we look at models of data structures and ask when does one model fit the data better than another.

The choice of one model of data structure over another may depend on how well the results can be interpreted within the paradigm of the subject matter that gave rise to the data. We thus beg the question of what is "structure" in the data and deal with the question of when does one model fit the data better than another. This is easier said than done in general in the field of data analysis and in cluster analysis in particular.

Although a solid, coherent foundation of cluster analysis and pattern recognition does not yet exist, a wide variety of different procedures have been applied to the problem of grouping data.

2.0 Applications of Cluster Analysis

We list here some of the diverse disciplines that have used clustering techniques:

> Geography--Regions in which various geographic or demographic characteristics are relatively similar are useful constructs for the geographer and have been found by clustering techniques. (Berry, 1960)
>
> Economics--For aggregation of various kinds of industries into groups having similar characteristics with regard to the economic analysis to be performed. (Fisher, 1958)
>
> Electrical Engineering--For the

detection of signals of unknown characteristics that recur frequently in a background of random noise. (Fralick, 1967)

Information Retrieval--To find classes of descriptors for articles and papers. (Dale and Dale, 1965)

Medicine--To group electrical cardiograms into subgroups. (Stark et al, 1962)

Numerical Taxonomy--To group species of living organisms into hierarchic trees by an explicit mathematically defined method to contrast the trees obtained with those obtained by older methods that contained considerable implicit judgement on the part of the taxonomist. (Sokal and Sneath, 1963)

Psychology and Sociology--To group people into types that may relate to treatment categories or behavior categories. (Tryon, 1967)

Statistics--For obtaining minimum variance stratified sampling partitions of a range of a variable. (Dalenius, 1951)

From this list it is clear that there has been a considerable interest in many disciplines in techniques that are able to take vectors and group them into subgroups in a way that makes some intuitive sense and that is useful in organizing data related to a variety of subject matter areas.

The need has always been around to group multivariate data, and some of the previous examples such as taxonomies of psychiatric groupings indicate that there was a hope that the groupings would have some organizing, some explanatory and some predictive power, for example, in taxonomy that we would be able to understand the evolutionary lineage of various organisms, or in the case of psychiatric groups to indicate treatment classes, where the same treatment could be used on a particular group.

The reader interested in pursuing further the work of cluster analysis in the field of psychiatry would be well advised to look at Katz, Cole, Barton (1968)---"The Role and Methodology of Classification in Psychiatry and Psychopathology", where he will find many points of view including those of clinicians as well as data analysts. In numerical taxonomy, the book by Sokal and Sneath (1965) remains the standard reference. In the area of market analysis the monograph by Green et al (1968) is one of the most complete we are aware of. The paper by Ball (1965) contains excellent references for the person interested to see the wide variety of different methods that were developed in numerous disciplines.

3.0 Clustering as a Data Analysis Tool

We see cluster analysis as one useful tool in the highly iterative process of data analysis. Clustering can be put to the following uses:

(a) It can suggest, by grouping the data into groups with high intra-group similarity, multiple working hypotheses that are appropriately vague and hence appropriately suggestive of alternate views of the data. It appears to us that this grouping is an important aspect of concept formation and of theory development, since in both cases grouping of things sufficiently similar to be considered as a unitary item can lead to important advances in understanding.

(b) Clustering provides a way of starting an analysis, and it can provide a flexible set of initial categories for further modification.

(c) It suggests possible ways of decomposing the data into simpler subsets of objects or variables that can be examined graphically, so that the analyst obtains a deeper understanding of the details of his data. (It appears that the interactive graphic computer may provide a useful tool that will allow the analyst to "zoom" from the details of his data up to a summary provided by a small number of cluster centers and the characteristics of the clusters.)

(d) Clustering can be used as a data manipulation algorithm. It can be used to reduce the dimensionality by first finding k cluster in a data space of p greater than k dimensions and from this obtaining the k-1 dimensional subspace of the original data "spanned" by these cluster centers. Whether other clustering algorithms provide a useful alternative to more traditional methods of principal components or factor analysis has yet to be shown.

(e) Clustering can be used to reduce the volume of a large data set by substituting for, say, 5,000 data samples, a non-random representative set of, say, 150 cluster centers that will tend to relatively adequately characterize the variation of the data. This similar set of cluster centers can then be used as input to such techniques as the multidimensional scaling programs that are not able to handle such large data volumes.

(f) In its loose role in exploring data, clustering can aid concept formation by providing a tractable description that still retains some of the subtleties of

the data that are lost when the data is described by a single mean and covariance matrix or other such simple descriptors. These techniques are useful for exploring the data and for examining relationships between the variables and how the particular data population chosen affects relationships between the variables. For example, the existence of two distinct subgroupings may seem to imply a correlation between two variables in a single population, when in fact it is primarily indicating a correlational relationship between the means of two different subpopulations. In the same way the choice of variables affects the groupings found, and this realization forces the analyst to reconsider the measurements he has chosen and sometimes the weightings applied to those measurements in a way that may not be as easy to overlook as it is with some other methods.

Clearly it is the digital computer that has made the use of clustering techniques attractive and in many instances possible for a variety of people. Not only has it provided the computational power to actually perform the clustering algorithms but it has also put us in the position where we can collect a great deal of multivariate data. The existence of this data has led us to want to understand, to explore, to search for patterns of association in the data, and this in turn has fed back into our needs for some techniques which could organize this data. This underlines the statement that convenience is not only dramatically helpful but it is also dramatically demanding. Once we have been given the tool, we find uses to which it can be put.

3.1 Clustering as Fitting

Clustering can be viewed as a classical goodness of fit problem. For example, one can assume a model for the data of a mixture of multivariate distributions and then test the fit of this model to the data. This approach has not proved very tractable in terms of computing, but some recent work of Wolfe (1967) shows promise.

More generally, one assumes a certain kind of structure and then attempts to fit this structure to the data. The analyst then examines departures from the assumptions, that is, the residuals, and indicates changes that need to be made in the model. In some instances it may be possible to indicate the directions that changes need to be made in order for the model to fit more satisfactorily. For example, consider principal components. If one fits a p-dimensional space with k less than p subspace, then the residuals for the individual data samples are well defined in that they are the distance between the data point in the subspace and the data point in the entire space.

In many types of clustering techniques it is not clear what the appropriate notion of goodness of fit is, and the notion of the residual needs to be further developed.

3.2 Evolution of Clustering Procedures and Criteria

Initially there was the search for algorithms that in some intuitively satisfying way found groups in the data. This work on formal clustering algorithms started as early as Tryon (1939) and is still going on, as exemplified by the ISODATA algorithm of Ball and Hall (1967), by the algorithms of Sokal and Sneath (1963), and by the k-means algorithms of MacQueen (1967). As these algorithms were used, there developed a concern as to the criteria being used and attempts to optimize criteria, as in the work of Ward (1963) and Singleton. The work of Hartigan (1967) is a further extension in which he formulates criteria for hierarchical structure. Then the focus shifted toward the development of algorithms to optimize the criteria found to be useful. This is exemplified by the $|T|/|W|$ criterion used by Friedman and Rubin (1967). The acceptance of this criterion shifted emphasis toward developing algorithms that would optimize that criterion.

Thus we see the need for clustering techniques expressed initially in a search for algorithms that would produce "clusters". By and large at this stage it was considered that exhaustive computation of all potentially useful partitions of the data was impossible, and the goal was to get one best answer as defined by the algorithm. Gradually, as awareness increased through the use of the algorithms, it became possible to see that criteria were in fact being optimized in many cases [see J. Gower (1967), where he shows that different algorithms lead to the same underlying structure]. As computational power increased sufficiently, it was, at times, possible to evaluate exhaustively the criteria, usually in the case of clustering for small problems. Finally the work then moved into developing algorithms dominated by a criterion where the focus now is on non-exhaustive search procedures that tend to optimize the criterion [see Friedman and Rubin (1967)].

It is clear that there is a large variety of techniques if one only looks at the algorithms used to cluster the data. But it is not equally clear that the groupings that result from using these algorithms will differ substantially. The implicit criterion controlling how groups are found that is defined by the very nature of the algorithm or by the measure of similarity that is used may not be substantially different from algorithm to algorithm. For example, the ISODATA algorithm (1967) and the Singleton-Kautz algorithm (Singleton, 1967) appear to be very different. And yet when, upon finding the groupings obtained to be very similar, we examined the algorithms further, we found that the Singleton-Kautz algorithm was explicitly attempting to find a minimum variance partition while the ISODATA algorithm was implicitly tending to find that same type of partitioning.

In Friedman and Rubin (1967) we see the beginnings of the comparison of different criteria

against the same data sets. In this paper the relation of some of the clustering methods with traditional multivariate statistical theory is elucidated. In this context a paper by Peter Ihm (1965) is very relevant.

In the area of factor analysis, a recent paper by M. Browne (1968) studies statistical properties of several methods of obtaining estimates from data. Again this paper is indicative of what is happening in cluster analysis. Different procedures were developed, and some amount of synthesis is take place, with people beginning to compare and evaluate different methods on data with known structures.

A striking example of the interaction of the methods of principal component analysis, factor analysis, cluster analysis and discriminant analysis applied to the analysis of psychiatric data is given by Friedman and Rubin (1968).

In the future, as computing power grows and changes in character with the development of the interactive graphic computer, it will be interesting to see if, in fact, algorithms are developed that are suitable for control by a man in a man-machine environment. In a much broader context, as people begin to develop on-line computer systems for data analysis, it is becoming increasingly apparent that one or more methods of cluster analysis will be included as data analysis tasks. It should be emphasized that in analyzing most large sets of multivariate data different methods of analysis will be employed on the same data set, and some form of cluster analysis will be most helpful in assessing the nature of the heterogeneity of the data as well as to group the data into more homogeneous groups for analysis. Further, as shown in Friedman and Rubin (1968), once having found some group structure, further multivariate analysis may be required to describe the data in a way that can be meaningfully interpreted in the paradigm of the subject matter of the data. The problem of interpreting the results of clustering procedures is very much a problem that still exists in the more traditional methods of multivariate analysis. We have a long way to go in our understanding of multivariate data.

4.0 Some Examples

In this section we consider three separate sets of data, each of which illustrates a particular aspect of using clustering techniques on real experimental data.

4.1 The Indian Studies Data

The desired goal of this clustering that used the Singleton-Kautz algorithm was to obtain a representative (non-random) sample of Indian cities that would allow a small number, say 10%, of the cities for detailed study as to the economic effects of improving their infrastructure, that is, their sewage system, electric power network and road structures. The aspect of this clustering that we would like to emphasize is shown in Figure 4, in which we have displayed the contribution of the sum of squared error of each of nine variables used in this particular clustering. (We clustered using a variety of sets of clustering variables. Figure 4 shows the result of one particular clustering.) The curves shown are essentially a "decomposition" of the total sum of squared error curve. For example, we see that in increasing the number of clusters from one cluster to two cluster that the contribution of Variables 1, 2, and 3 to the total sum of squared error markedly decreased with most other variables not being substantially decreased in their sum of squared error. In increasing from two to three clusters we see that Variables 6 and 9 cooperated, and in going from three to four clusters we see that Variable 5 contributed markedly, with Variable 1 contributing somewhat. This ability to perceive the way in which variables interact in reducing the sum of the squared errors suggests that strong relationships do exist between these variables. With this information, it is then possible to go back and examine the details of that interrelationship for just those three variables, using clustering or perhaps even graphic techniques. (This desire to go back and forth has been a major factor in the development of PROMENADE--a multivariate data analysis system that uses interactive computer graphics. See Ball and Hall, 1967.)

These relationships between variables and the observation that pairs or triples of variables tended to cooperate suggested to us that we ought to re-examine the relationship between principal components analysis and minimum variance clustering partitions. This comparison is still in process and has not yet been completed.

4.2 Job Satisfaction Data

This data consisted of 209 responses by Air Force scientists to questions that sought to ascertain the degree to which they were satisfied with their present employment and to elicit some of the factors related to satisfaction. One useful grouping consisted of seven clusters. We now describe the average response patterns for two of these clusters in order to show the suggestiveness of the groupings in suggesting explanations for job satisfaction. The work profile for Group 7 describes the people of this group as being between 35 and 50 years of age (95%); highly educated (50% Ph.D., 50% M.S.); civilians primarily performing basic research (80%); planning to work for the employing organization in ten years; publishing professional papers frequently (75% have five or more in the last five years); quite satisfied with their jobs (55% highly satisfied and only 15% dissatisfied). This can be compared to the work profile of Group 6, where 20% are under 30 years of age, have a moderate degree of education (33% M.S., 67% B.S.); are civilians; primarily performing applied or basic applied research (60%); planning to be working in the same organization in ten years; publishing infrequently (45% have zero to two publications in the last five years); only moderately satisfied with their jobs (30% highly satisfied and 35% dissatisfied). Another grouping found was that of the young researcher who had recently obtained his

Ph.D. and was now fulfilling his military obligation working in an Air Force laboratory at relatively low pay compared to civilians working in the same capacity. This group tended to be highly dissatisfied.

This data set indicated that the high interrelationship between the various questions on this questionnaire did almost necessarily insure that a useful grouping would result in the clustering and that the profiles would allow for interpretation in a convenient fashion. This coherence of the variation measured by the questions on the questionnaire in this data set did not exist in the data set which we discuss next.

4.3 Controversial Issues Data

This data consisted of questions asked about attitudes on various controversial subjects, including questions about religion, sex, politics, and other personal attitudes. The questionnaire was potentially designed to cut across many different attitudes, and there was far less interaction between various questions on the data set.

We clustered the data using all of the questions and were unable to discover any useful or interesting clusterings. We then focused on one set of questions relating to attitudes about sex. Out of this focusing on about one quarter of the questions, we were able to obtain extremely well-defined clustering that had a simple and straightforward explanation.

This data set led us to conclude that there is a great need for methods that search for variables on which to do the clustering in addition to the already existing techniques such as principal components. In a broader sense, the results suggest that there is a kind of interaction between the variables used and the objects used that suggests iterative procedures that successively define subpopulations both of objects and variables that lend themselves to interpretation. The task would then become one of interrelating results obtained on the different subpopulations--perhaps by using the subcategorizations obtained from considering just subsets of the variables and and subpopulations of objects.

5.0 Summary Remarks

Computers have added conceptual as well as algorithmic dimensions to statistics, and clustering techniques are one evidence of this change.

The movement caused in statistics by the use of the computer has led to new flexibility in criteria and the acceptance of more empirical methods. For example, other criteria than least squares are now acceptable. This movement makes it easier to accept clustering techniques as part of the statistics tradition.

Clustering itself is in transition. Movement is taking place from a loosely structured body of ad hoc algorithms toward a coherent set of tools whose interrelationships with each other and with other existing multivariate techniques are becoming known.

References on Clustering

1. Ball, Geoffrey H. (1965), "Data Analysis in the Social Sciences", American Federation of Information Processing Societies Conference Proceedings, Fall Joint Computer Conference, Vol. 27, Part 1, Washington: Sparton Books, London: Macmillan, pp. 533-560.

2. Ball, Geoffrey H. and David J. Hall (1967), "A Clustering Technique for Summarizing Multivariate Data", Behavioral Sciences, 12, No. 2, pp. 153-155.

3. Ball, Geoffrey H. and David J. Hall (1967), "PROMENADE, An On-Line Pattern Recognition System", Stanford Research Institute Technical Report No. RADC-TR-67-310, Menlo Park, California, pp. 1-126.

4. Berry, Brian J.L. (1960), "An Inductive Approach to the Regionalization of Economic Development", #62, University of Chicago, pp. 78-107.

5. Bonner, R.E. (1964), "On Some Clustering Techniques", IBM Journal, 22, pp. 22-32.

6. Browne, Michael (1968), "A Comparison of Factor Analytic Techniques", Psychometrika, Vol. 33, No. 3.

7. Dale, A.G. and N. Dale (1965), "Some Clumping Experiments for Associative Document Retrieval", Am. Documentation, Vol. 16, No. 1, pp. 5-9.

8. Dalenius, Tore and M. Burney (1951), "The Problem of Optimum Stratification II", Skandinavisk Aktuarietidskrift, Vol. 34, pp. 133-148.

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

10. Fisher, Walter D. (1965), "Constructive Partition and Aggregation", Kansas State University, Manhattan, Kansas, pp. 1-19.

11. Forgy, Edward W. (1963), "Detecting 'Natural' Clusters of Individuals", Department of Psychiatry, University of California Medical Center, Los Angeles, pp. 1-10.

12. Fralick, S.C. (1967), "Learning to Recognize Patterns without a Teacher", IEEE Transactions on Information Theory, IT-13, No. 1, pp. 57-64.

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

14. Friedman, H.P. and J. Rubin (1968), Chapter 5 "Logic of Statistical Procedures", The

Borderline Syndrome, by Grinker et al., Basic Books.

15. Gower, J.C. (1967), "A Comparison of Some Methods of Cluster Analysis", Biometrics.

16. Green, P. et al. (1968), "Analysis of Marketing Behavior Using Nonmetric Scaling and Related Techniques", Technical Report, Marketing Science Institute, University of Pennsylvania.

17. Guttman, Louis (to be published December 1968), "A General Non-Metric Technique for Finding the Smallest Coordinate Space for a Configuration of Points", Psychometrika.

18. Hartigan, J.A. (1967), "Representation of Similarity Matrices by Trees", Journal of the American Statistical Association, Vol. 62.

19. Ihm, P. (1965), "Automatic Classification in Anthropology", The Use of Computers in Anthropology, Edited by Dell Hymes, London: Mouton and Company.

20. Johnson, Stephan C. (1967), "Hierarchical Clustering Schemes", Psychometrika, 32, No. 3, pp. 241-254.

21. Kruskal, J. (1964), "Multidimensional Scaling by Optimizing Goodness of Fit to a Non-Metric Hypothesis", Psychometrika, 29, pp. 1-28.

22. Kruskal, J.B. (1964), "Nonmetric Multidimensional Scaling: A Numerical Method", Psychometrika, 29, No. 2, pp. 115-129.

23. MacQueen, J.(1967),"Some Methods for Classification and Analysis of Multivariate Observations", 5th Berkeley Symposium on Mathematics, Statistics and Probability, 1. pp. 281-297.

24. Mattson, R.L. and J.E. Damman (1965), "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems", IBM Journal, pp. 294-302.

25. Medgyessy, Pal (1961), Decomposition of Superpositions of Distribution Functions, Publishing House of Hungarian Academy of Science, Budapest.

26. Nunnally, Jim (1962), "The Analysis of Profile Data", Psychological Bulletin, 59, No. 4, pp. 311-319.

27. Rao, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research, Sankhya, Series A, Vol. 26, Part 4, pp. 329-358.

28. Rogers, David J. and Taffee T. Tanimoto (1960), "A Computer Program for Classifying Plants", Science, 132, pp. 1115-1118.

29. Rubin, J. (1967), "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem", Journal of Theoretical Biology, 15, pp. 103-144.

30. Sammon, John W.,Jr. (1968), "An Adaptive Technique for Multiple Signal Detection and Identification", Pattern Recognition, Edited by L. Kanal, Thompson Book Company, Washington, D.C., pp. 409-439.

31. Sebestyen, George S. (1966), "Automatic Off-Line Multivariate Data Analysis", Proceedings of Fall Joint Computer Conference, Spartan Books, pp. 685-694.

32. Shepard, R.N. (1962), "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function I, Psychometrika, 27, pp. 125-140.

33. Singleton, R.C. (1967), Private Communications, Stanford Research Institute, Menlo Park, Calif.

34. Sokal, R.R. and P.H.A. Sneath (1963), Principles of Numerical Taxonomy, W.H. Freeman and Company, San Francisco.

35. Stanat, Donald F. (1968), "Unsupervised Learning of Mixtures of Probability Functions", Pattern Recognition, Edited by L. Kanal, Thompson Book Company, Washington, D.C., pp. 357-389.

36. Stark, Lawrence, Mitsuharu Okajima and Gerald M. Whipple (1962), "Computer Pattern Recognition Techniques: Electrocardiographic Diagnosis", Communications of the Association for Computer Machinery, 6, No. 10, pp. 527-532.

37. Torgerson, W. (1968), "Multidimensional Representation of Similarity Structures", The Role and Methodology of Classification in Psychiatry and Psychopathology, Edited by Katz, Cole and Barten, U.S. Department of Health, Education and Welfare.

38. Tryon, Robert C. (1967), "Person Clusters on Intellectual Abilities and on MMPI Attributes", Multivariate Behavioral Research, Vol. 2, pp. 5-34.

39. Tucker, Ledyard R. (1968), "Cluster Analysis and the Search for Structure Underlying Individual Differences in Psychological Phenomena", Conference on Cluster Analysis of Multivariate Data, New Orleans, Louisiana, pp. 10.01-10.17.

40. Wallace, David (1968), "Cluster Analysis", International Encyclopedia of the Social Sciences, Cromwell Collier, pp. 519-524.

41. Ward, Joe H., Jr. and Marion E. Hook (1963), Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles", Educational and Psychological Measurement, 23, No. 1.

42. Wolfe, John H. (1967), "NORMIX--Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions", Technical Report, U.S. Naval Personnel Research Activity, San Diego, California, pp. 1-31.