Journal of the
Military Operations Research
Society of Korea, Vol. 6, No. 2
December, 1980

Clustering Technique for Multivariate Data Analysis

Lee Jin Ki *

ABSTRACT

The multivariate analysis techniques of cluster analysis are examined in this article. The theory and applications of the techniques and computer software concerning these techniques are discussed and sample jobs are included.

A hierarchical cluster analysis algorithm, available in the IMSL software package, is applied to a set of data extracted from a group of subjects for the purpose of partitioning a collection of 26 attributes of a weapon system into six clusters of superattributes.

A nonhierarchical clustering procedure were applied to a collection of data of tanks considering of twenty-four observations of ten attributes of tanks. The cluster analysis shows that the tanks cluster somewhat naturally by nationality. The principal componant analysis and the discriminant analysis show that tank weight is the single most important discriminator

*) Office of OR/SA, MND

among nationality although they are not shown in this article because of the space restriction.

This is a part of thesis for master's degree in operations research.

1. INTRODUCTION

A. ORIGIN AND THEORY

Clustering is the grouping of similar objects. The principal functions of clustering are to name, to display, to summarize, to predict, and to aid in interpretation of data with many dimensions. Clustering techniques were first developed in the field of biological taxonomy. It is one of several methodologies included in the broader category called classification.

The cluster analysis problem is the last step we consider in the progression of category sorting problems. While in discriminant analysis some part of the structure is known and missing information is estimated from labeled samples, the operational objectives of clustering is to classify new observations, that is, recognize them as members of one category or another. In cluster analysis little or nothing is known about the category structure. All that is available is a collection of observations whose category membership are known. We seek to discover a category structure which fits the observations.

The observation can be written as the usual n p data matrix.

$$x = \begin{bmatrix} x_{11} & \cdots & x_{ip} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} X_1' \\ \vdots \\ X_p' \end{bmatrix} \qquad (1)$$

The problem may be stated as one of finding the "natural groups," which means to sort the observations into groups such that the degree of "natural association" is high among members of the same group and low between members of different groups.

Cluster analysis techniques have been applied in many fields of study. The literature is both volumnious and diverse, the terminology differing from one field to another. "Numerical taxonomy" is frequently substituted for cluster analysis among biologists, botanists, and ecologists, while some social scientists may refer "typology." Other frequently encountered terms are pattern recognition and partitioning. While discriminant analysis has been studied by statisticians for nearly 45 years, cluster analysis has only recently come to statistical notice. Any method which partition a set of objects into subsets on the basis of measurements taken on every object qualifies as a clustering method.

Most of the well known clustering techniques fall into one of two main categories: (a) hierachical and (b) nonhierachical (partitioning). The former is one in which every cluster obtained at any stage is a merger of clusters at previous stages. The

nonhierachial procedures however form new clusters by lumping and splitting old ones. We consider both categories shortly.

In a geometric sense, every observation may be viewed as a point in p-dimensional Euclidean space. This swarm of data points may contain dense regions or "clouds" of data points which are separable from other regions containing a low density of points. These denser regions constitute what are known as clusters. In one and two dimensional cases, it is easy to visualize and to detect the clusters from scatter plots, assuming that the clusters exist. In higher dimensions, clustering becomes extremely difficult without the aid of a computer.

Mathematical clustering techniques usually require a measure of similarity to be defined for every pairwise combination of the entities to be clustered. In order to solve the cluster problem, it is desirable to define the terms "similarity" and "difference" in a quantitative fashion. A researcher would assign two observations to the same group if the distance between them is sufficiently small, or to different clusters if this distance is sufficiently large.

At this point, two questions may be brought on. The first one is "how do we measure the distance between the observations?" and the second one is "how small is small enough?" and how large is large enough? These will be discussed in the following sections.

B. MEASURES OF DISTANCE

(1) General

Let E_p be a symbolic representation for a measurement in p-dimensional space and let X,Y, and Z be any of these points in E_p. Then any nonnegative real-valued function D(X,Y) satisfying the following conditions qualifies as a distance function (or metric).

(a) $D(X,Y) = 0$ if and only if $X = Y$

(b) $D(X,Y) \geqq 0$ for all X and Y in E_p

(c) $D(X,Y) = D(Y,X)$

(d) $D(X,Y) \leqq D(X,Z) + D(Y,Z)$

Many clustering algorithm assume such distances given and set about constructing clusters of objects within which the distances are small. The choice of distance function is no less important than the choice of variables to be used in the study. A serious difficulty in choosing a distance lies in the fact that a clustering structure is more primitive than a distance function and that knowledge of clusters changes the choice of distance function. Thus a variable that distinguishes well between two established clusters should be given more weight in computing distances than a "junk" variable that distinguishes badly.

(2) Euclidean Distance

The Euclidean distance between the I-th and K-th observations of a data matrix X is defined as

$$D(I,K)=\left[\sum_{1\leq J\leq p}\{X(I,J)-X(K,J)\}^2\right]^{1/2} \qquad (2)$$

where J is J-th variable. In one, two, or three dimensional space, this is just a "straight line" distance between the vectors corresponding to the I-th and K-th observations. When the variables are measured in different units, it is necessary to prescale the variabes to make their values comparable, or, equivalently, to compute a weighted Euclidean distance.

$$D(I,K)=\left[\sum_{1\leq J\leq p}W(J)(X(I,J)-(X(K,J))^2\right]^{1/2} \qquad (3)$$

This form of distance is not necessary if all variables are measured on the same scale. However, even in this case, weights might be used to increase or decrease the importance of same variable. Various weighting schemes have been utilized in practice. One common weighting scheme lets W(J) be the reciprocal of the variance of variable J.

A general class of squared distance functions is provided by utilizing positive definite quadratic forms. Specifically, if p represents a P-dimensional observation to be assigned to one of s groups, then to measure the squared distance between the observation β and the centroid (mean vector) of the i-th group one may consider the function

$$D_i = (\beta - x_{i.})^T M (\beta - x_{i.}) \qquad (4)$$

where M is a positive definite matrix to ensure that $D_i \leqslant 0$. Different distance functions are represented by different choices of the matrix M. When M = I (the identify matrix) the resulting metric is the standard Euclidean distance. Distances with the Euclidean metric are shown in Figure 1a. The variance within the data may make the unweighted Euclidean metric inappropriate. As shown on the Figure 1b, where X has a larger variance than Y, one may wish to weight a deviation in the X direction less than an equal deviation in the Y direction. This is a weighted Euclidean distance frunction which makes point A and B equidistance from the origin. In this case, the matrix M is diagonal elements which are the reciprocals of the variances of the different variables.

Extending this idea further, it may be possible to consider the covariance among variables as well. Figure 1c shows how the axis may be rotated so that the major axis is oriented in a direction of reflecting the positive correlation between X and Y. Again, points on the same ellipse are considered equidistance from the origin. The matrix M in this case is the inverse of the covariance matrix.

Further extension of this concept will explain some sort of generalized distance function. If C_i represents the covariance matrix of the i-th cluster then the distance function

$$D_i = (B - \bar{x}_{i.})^T \; C_i^{-1} \; (B - \bar{x}_{i.}) \qquad (5)$$

uses the appropriate covariance structure when determining the distance to a particular cluster centroid. Since C_i changes to reflect the dispersion internal to each particular cluster, the use of this metric exploits differences in the dispersion characteristics of the different groups. As shwon on Figure 1d, not how a new observation (denoted by u) is closer to the centroid of group one (G1) in terms of Euclidean distance but is more likely to be assigned to group two (G2) when using the C_i matrix.

(3) Mahalanobis Distance

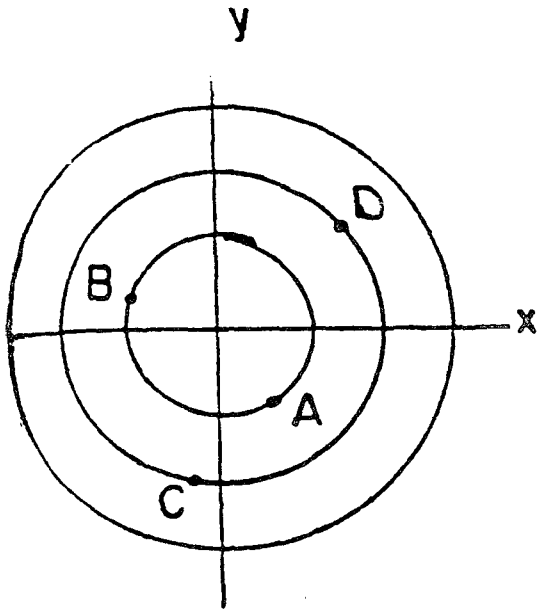
Another choice for the M matrix in equation (4) is p^{-1} where p represents the pooled within groups covariance matrix of all the clusters.

$$P = \frac{1}{\sum_{i=1}^{G} (n_i - 1)} W \qquad (6)$$

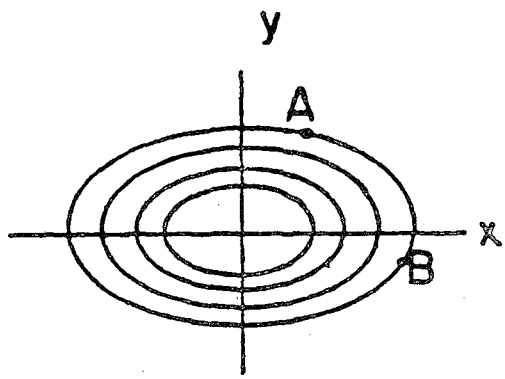
where

$$W = \sum_{k=1}^{G} W_k$$

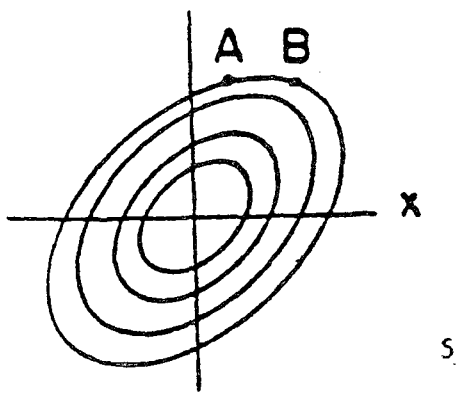
This distance is the well known Mahalanobis distance. Note that P does not change from group to group. The ensure the non-singularity of P it must be true that $p \leq (N - G)$, where N represents the total number of observations over all groups. Rewriting the distance,



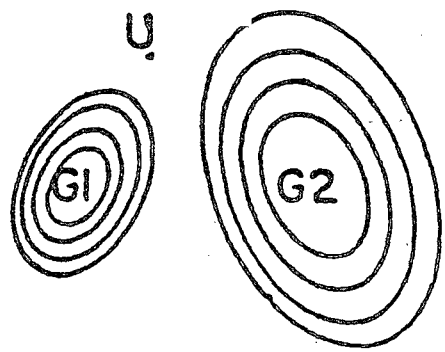
1a. Euclidean measure of squsred distance.



1b. Measure of squared distance with different weight for variab.



1c. Generalized squared distance measure.



1d. Classlfication when within-group dispersions are different.

Figure 1. Euclidean Distance

$$D_i = (B - \overline{x}_{i.})^T W^{-1} (B - \overline{x}_{i.}) \qquad (7)$$

defines a distance between mean vectors β and $\overline{x}_{i.}$ and common covariance matrix W. The Mahalanobis distance function adjusts for both scale of measurement of the variables and covariation among the variables. Use of this metric is equivalent to computing distances on variables transformed to their principal components. This metric is invariant under any nonsingular transformation of original variables. For consider the transformation

$$Y = BX \qquad (8)$$

and let $D(Y_i,Y_j)$ represent Mahalanobis distance between Y_i and Y_j.

$$\begin{aligned} D(Y_i,Y_j) &= (Y_i - Y_j)^T P_Y^{-1}(Y_i - Y_j) \\ &= (BX_i - BX_j)^T P_Y^{-1}(BX_i - BX_j) \\ &= (X_i - X_j)^T B^T P_Y^{-1} B(X_i - X_j) \\ &= (X_i - X_j)^T B^T (BP_XB^T)^{-1}B(X_i - X_j) \\ &= (X_i - X_j)^T P_X^{-1} (X_i - X_j) \\ &= D(X_i,X_j) \end{aligned}$$

Some other common metrics are listed below:

(a) L_1 norm (City Block)

$$D(X_i,X_j) = \sum_{k=1}^{p} | X_{ki} - X_{kj} |$$

(b) L_p norm (Minkowsky Metrics)

$$D(X_i,X_j) = (\sum_{k=1}^{p} | X_{ki} - X_{kj} |^p)^{1/p}$$

3. Uniform norm

$$D(X_i,X_j) = \underset{k = 1, 2, \ldots,p}{\text{Superemum}} \{ | X_{ki} - X_{kj} | \} \qquad (9)$$

C. HIERARCHICAL CLUSTERING

(1) General

The previously discussed distance measures may be used to construct a similarity matrix describing the length of all pairwise relationships among the entities (variables or data units) in the data set. The methods of hierarchical cluster analysis operate on this similarity matrix to construct a tree depicting specified relationships among the entities. As shown on Figure 2, the branches on the left each represent one entity while the root represents the entire collection of entities. Moving down the tree from the branches toward the root depicts increasing aggregation of the entities into clusters. Hierarchical clustering methods which build a tree from branches to root often are called agglomerative methods.

Once a tree is constructed for N entities, the analyst may choose from as many as N sets of clusters. These clusters are nested. From the agglomerative view, when two entities are merged they are joined together permanently and considered as one entity for later merges; from the divisive view, when a group of entities is split into two parts, the parts are separated permanently and may be treated independently for the remainder of the analysis.

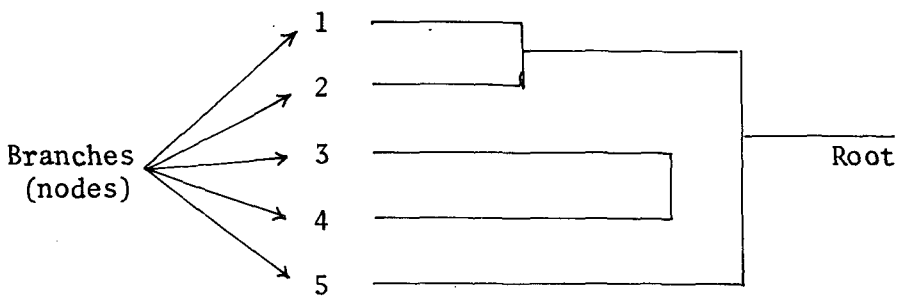


Figure 2. Tree for Hierarchical Clustering

Herein lie both the strength and weakness of hierarchical methods: by taking early decisions as permanent, the number of possibilities that need be examined is reduced greatly as compared with complete enumeration; but this same convention precludes discovering early mistakes or capitalizing on later opportunities.

There are three major hierarchical clustering concepts :

(a) Linkage Methods

(b) Centroid Methods

(c) Error sum of squares or variance methods.

All of these methods are suitable for clustering data units. However, only the linkage methods are considered in this research.

(2) The General Agglomerative Procedure

Let S_{ij} be the similarity between entities i and j as defined by one of the distance measures previously discussed. Assuming that the similarity is symmetric, the complete schedule of similarities for all $\binom{N}{2} = \frac{1}{2}N\ (N - 1)$ possible pairwise combinations of entities may be arrayed in a lower triangular similarity matrix as in Figure 3. The s_{ij} entries are nonnegative. This limitation is of consequence only for correlation and the cosine of the angle between vectors; the distinction between positive and negative association cannot be utilized in these clustering methods.

$$S = \begin{vmatrix} s_{21} & & & & \\ s_{31} & s_{32} & & & \\ s_{41} & s_{42} & s_{43} & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ s_{n1} & s_{n2} & s_{n3} & \cdots & s_{n(n-1)} \end{vmatrix}$$

Figure 3. Lower Triangular Similarity Matrix

A simple remedy is to use the absolute value or the square of the measure if it can assume negative values. Once the matrix is defined, the process of clustering entities is almost trivially simple. The general procedure for agglomerative clustering on a

data matrix is as follows :

(a) Begin with n clusters each consisting of exactly one entity. Let the clusters are labled with the numbers 1 through N.

(b) Search the similarity matrix for the most similar pair of clusters. Let the chosen clusters be labeled p and q and let their associated similarity be s_{pq}, $p > q$.

(c) Reduce the number of clusters by 1 through merger of clusters p and q. Label the product of the merger q and update the similarity matrix entities in order to reflect the revised similarities between cluster q and all other existing clusters. Delete the row and column of S pertaining to cluster p.

(d) Perform steps b and c a total of N-1 times (at which point all entities will be one cluster). At each step record the identity of the clusters which are merged and the value of similarity between them in order to have a complete record of the results.

Different agglomerative methods are implemented by varying the procedures used for defining the most similar pair at step b and for updating the revised similarity matrix at step C. The similarity matrix is a given array of numbers. The numerical execution of the clustering procedures is completely independent of how the similarity values were generated or whether the entities to be clustered are variables or data units. However, it is necessary to

make a direct distinction between distance-like measures (the smallest values correspond to the most similar pairs) and correlation-like measures (the largest values correspond to the most similar pairs); the essential difference is whether the search for the most similar pair involves seeking the minimum or maximum entry in the similarity matrix.

(3) Single Linkage

The method of single-linkage cluster analysis is the simplest of all hierarchical techniques. At each stage, after clusters p and q have been merged, the similarity between the cluster (labeled t) and some other r is determined as follows:

(a) If s_{ij} is the distance-line measure

$$s_{tr} = \min (s_{pr}, s_{qr}) \tag{10}$$

The quantity s_{tr} is the distance between the two closest members of clusters t and r. If clusters t and r were to be merged, then for any entity in the resulting cluster the distance to its nearest neighbor would be at most s_{tr}.

(b) If s_{ij} is a correlation-like measure

$$s_{tr} = \max (s_{pr}, s_{qr}) \tag{11}$$

The quantity s_{tr} is the similarity between the two most similar entities in clusters t and r. If clusters t and r were to be merged, then for any entity in the resulting cluster there would be

at least one other entity in the same cluster such that the pair would have a similarity at least as large as s_{tr}.

The method is known as single linkage because clusters are joined at each stage by the single shortest or strongest link between them. Since the updating process involves choosing only the minimum or maximum single-linkage clustering is invariant to any transformation which leaves the ordering of the similarities unchanged; that is, any monotonic transformation.

(4) Complete Linkage

The complete-linkage method is related to the single-linkage method and is no more difficult to execute. At each stage, after clusters p and q have been merged, the similarity between the new cluster (labeled t) and some other cluster r is determined as follows:

(a) If s_{ij} is distance-like measure

$$s_{tr} = \max (s_{pr}, s_{qr}) \tag{12}$$

The quantity s_{tr} is the distance between the most distant members of clusters t and r. If clusters t and r were merged, then every entity in the resulting cluster would be no farther than s_{tr} from every other entity in the cluster. The value of s_{tr} is the diameter of the smallest sphere which can enclose the cluster resulting from the merger of clusters t and r.

(b) If s_{ij} is a correlation-like measure

$$s_{tr} = \min (s_{pr}, s_{qr}) \tag{13}$$

The quantity s_{tr} is the similarity between the two most dissimilar entities in clusters t and r. If clusters t and r were to be merged, then every entity in the resulting cluster would have a similarity of at least s_{tr} with every other entity in the cluster.

The method is called complete linkage because all entities in a cluster are linked to each other at some maximum distance or minimum similarity. Such a cluster is called a "maximally connected subgraph" in graph theory. In contrast to the signle-linkage method, interpretation of the clusters can be made only in terms of the relationships within individual clusters; there is no particularly useful interpretation involving the differences between clusters. Like the single-linkage method, complete-linkage cluster analysis is invariant to monotonic transformations of the similarity measure. Jonson (1967) discusses this property in both single and complete linkage methods.

D. NONHIERARCHICAL CLUSTERING

Nonhierarchical clustering methods are designed to cluster data units into a single classification of g clusters, where g either is specified a priori or is determined as a part of the clustering method. The central idea in most of these methods is

to choose some initial partition of the data units and then alter cluster memberships so as to obtain a better partition. The various algorithms which have been proposed differ as to what constitutes a "better partition" and what methods may be used for achieving improvements.

The broad concept for these methods is very similar to that underlying the steepest descent algorithms used for unconstrained optimization in nonlinear programming. Such algorithms begin with an initial point and then converge to a local optimum, moving one step at a time, the value of the objective function improving at each step.

The methods of nonhierarchical clustering typically, may be used with much larger problems than the hierarchical methods because it is not necessary to calculate and store the similarity matrix; it is not even necessary to store the data set. In general, the data units are processed serially and can be read from tape or disk as needed. This caharacteristic makes it possible, at least in principle, to cluster arbitrary large collections of data units.

In this research, we consider only the partitioning method known as "K-MEANS" which was developed by MacQueen (15). He used the term "K-MEANS" to denote the process of assigning each data unit to that cluster (of k clusters) with the nearest centroid (mean vector). The cluster centroid changes with each transfer of an observaiton.

The decomposition of the total scatter matrix into within and between groups matrices suggests possible optimality criteria to be

used in a clustering algorithm. One would like the within-groups scatter to be small relative to the between-groups scatter. Various trial clusterings could be formed using the W and B matrices as a basis for the optimality criteria which determine the best clustering. A possible choice for a criterion is to minimize trace W over all partitions into g groups. Since T is constant over all partitions, minimizing trace W is equvalent to maximizing traces B since

$$\text{trace } T = \text{trace } W + \text{trace } B \qquad (14)$$

Although trace W is invariant under an orthogonal transformation, it is not invariant under other non-singular linear transformations.

McRae (16) points out that trace W equals the total within group sum of squares, hence the "minimum variance partition" cluster solution is found by minimizing trace W .

Considerable study has been developed to alternative criteria such as those based on multivariate statistical analysis techniques, especially the methods of linear discriminant analysis and multivariate analysis of variance. Assuming the p variables are not linearly dependent, then as long as $p = N - g$, W is positive definite symmetric and so is W^{-1}. Attempts to make B and W as different as possible lead one to solving the determinantal equation:

$$|B - \lambda W| = 0 \qquad (15)$$

The solutions λ_i are the eigenvalues of the matrix $W^{-1}B$ as in discriminant analysis. There are t non-zero eigenvalues, where t is the minimum of p and g-1. This is a consequence of the fact that, if g is less than p, the g group means are considered in a (g-1)-dimensional hyperplane. When g = 2 the analysis is equivalent to two-group discriminant analysis. Linear discriminant analysis would take the vectors originally described in p-dimensional coordinate system and transform the basis to a t-dimensional system. Maximizing the largest of these eigenvalues is a criterion suggested by S.N. Roy and maximizing the trace of $W^{-1}B$, however is a criterion suggested by Hotelling. In both cases, large values for these statistics are sought in clustering algorithms since large values indicate large differences among (between) groups. Minimizing the ratio of determinants $|W| \div |T|$ is a criterion widely known as Wilks' lambda discussed in the discriminant analysis. Since T is the same for all partitions, this criterion is equivalent to minimizing determinant W . Both trace $W^{-1}B$ and $|T| \div |W|$ may be expressed in terms of the eignevalues of $W^{-1}B$.

$$\left| \frac{T}{W} \right| = \prod_{i=1}^{t} (1 + \lambda_i) \qquad (16)$$

$$\text{trace} \quad W^{-1}B = \sum_{i=1}^{t} \lambda_i \qquad (17)$$

where $t = \min(p, g-1)$. Therefore minimizing det W is equivalent to maximizing $\Pi(1 + \lambda_i)$.

Friedman and Rubin (6) describe the advantages of the various criteria. Those based on multivariate statistical considerations (all but trace W) are invariant under changes in scale for varibles (non-singular linear transformation). In fact, they are the only invariants for W and B under such transformations. In addition, the multivariate criteria may take into account covariation among the variables.

2. ANALYSIS OF MULTIVARIATE UTILITY DATA

To illustrate hierarchical clustering we applied the technique described in the previous chapter to partition a set of twenty six attributes of a close-air support weapon system into a smaller collection of "superattributes". As part of an effort to evaluate the military utility of a proposed alternative U.S. Marine Corps air support rada system, AN-TPQ/27. Barr and Richards (4) extracted 26 attributes of the TPQ-27 and a baseline system, the AN-TPQ/10, and then had members of the Operational Test and Evaluation Team assess the utility of the TPQ/27 relative to that of the TPQ/10. In order that the additive model used to combine unidimensional relative utilities into a system relative utility be justifiable, it is necessary that the utilities satisfy certain independence properties described in Keeney and Raiffa (12).

Because those independence properties are very difficult for decision makers to verify for complex alternatives like the weapon

systems under study, Professors Barr and Richards attempted instead to work with the attributes to try to generate a new collection which would likely satisfy, at least approximately, the conditions required to justify the additive model.

The original collection of 26 attributes is as follows:

(1) Portability
(2) Durability
(3) Time to Set Up
(4) Time to Take Down
(5) Ease of Assigning Aircraft to Targets
(6) Number of Aircraft Controlled
(7) Number of Targets
(8) Communications
(9) Mission Flexibility
(10) ASRT Survivability
(11) Time to Locate and Acquire Aircraft
(12) Accuracy of Tracking
(13) Accuracy of Delivery
(14) Range
(15) Aircraft Vulnerability
(16) Aircraft Attack Throughout
(17) Base of Adjustment and Evaluation of Results
(18) Accuracy of Feedback
(19) Ease of Operation
(20) Man-Machine Compatibility
(21) Training Requirements
(22) Reliability
(23) Maintainability
(24) Supportability
(25) Availability
(26) Documentation

Table I. Data Matrix

	1	2	3	4	5	6	7	8	9	10	11	12
1	6	2	1	3	2	4	3	1	3	1	1	1
2	1	2	2	3	1	3	2	1	1	5	1	2
3	6	1	1	5	2	4	3	1	6	1	2	1
4	6	1	1	5	2	5	3	1	6	1	2	1
5	2	5	3	1	3	1	1	2	2	2	3	3
6	2	7	4	1	4	1	1	2	4	2	3	3
7	2	7	4	1	4	1	1	2	4	2	3	3
8	3	5	4	7	5	6	1	3	1	3	3	6
9	2	5	4	1	3	1	1	2	4	2	3	3
10	9	6	5	6	5	7	8	3	7	3	2	4
11	2	8	6	4	3	2	1	4	4	2	3	3
12	3	8	7	4	4	2	6	5	4	6	4	3
13	3	8	7	4	4	2	6	5	4	6	4	3
14	3	8	4	4	4	2	6	5	4	2	4	3
15	7	6	5	6	5	7	7	3	7	3	7	4
16	8	8	6	1	4	1	7	4	4	2	7	3
17	8	8	8	4	3	2	5	6	5	4	5	3
18	4	8	8	4	6	2	5	6	5	6	5	3
19	4	5	3	1	3	1	1	2	2	4	3	3
20	4	5	3	3	3	3	3	1	2	4	8	3
21	5	4	9	2	3	8	4	7	2	4	6	5
22	1	3	2	3	1	3	2	7	1	5	6	2
23	1	3	9	3	1	3	2	7	1	5	6	2
24	1	3	4	3	1	3	2	7	3	5	6	2
25	1	3	2	3	1	3	2	7	1	5	6	2
26	5	4	9	2	6	8	4	7	2	4	6	5

It is clear from observing the above collection that some of the attributes are highly correlated and nonredundant. If one tries to assign an importance weights to each attributes separately, there is a distinct likelihood that some of the overlapping strongly into related attributes might effectively be double or triple weighted or more producing biased result. It is an effort to prevent this from happening, Barr and Richards aksed the utility assessment team to partition the 26 attributes into a smaller collection in such a way that attributes within a group are similiar and attributes in different groups are unrelated the sense that utility assessments for attributes in one group do not depend on the amounts of attributes in any other group.

The total number of groups was not prespecified. Instead, each team member was allowed to partition the 26 attributes into any number of groups. The resulting multivariate data array is show in Table I. An element x_{ij} is the number of the group into each team member j put attribute i.

Let us define a distance measure for this data array as follows:

$$D(a_i, a_k) = \sum_{j=1}^{12} (1 - I(x_{ij}, x_{kj})) \qquad (18)$$

where a_i represents an attribute i and

$$I(x_{ij}, x_{kj}) = \begin{cases} 1 & \text{if } x_{ij} = x_{kj} \\ 0 & \text{if } x_{ij} \neq x_{kj} \end{cases} \qquad (19)$$

It is easy to verify that D is a metric as defined in Chapter 1. Since we will actually work with a similarity measure in the hierarchical cluster procedure, we define the similarity between two attributes a_i and a_k as

$$S(a_i , a_k) = \sum_{j=1}^{12} I(x_{ij} , x_{kj}) \qquad (20)$$

One can see from this definition that the similarity between two attributes a_i and a_k is simply the number of team members who placed attributes a_i and a_k in the same partition. For example,

$$S(a_1, a_2) = 0 + 1 + 0 + 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 1 + 0$$
$$= 4$$

Either S or D can be used in the computer program IMSL. One need only indicate whether he wants a correlation-like (larger values imply more similar) measure or a distance-like measure (smaller values imply more similar). We selected to use the former method. The similarity matrix extracted from the data is shown in Table II. We present only lower triangular elements since $S(a_i, a_i) = 12$ for all i and the matrix is symmetric; i.e., $S(a_i, a_k) = S(a_k, a_i)$. Zero values are not written.

The results from the hierarchical clustering are shown in Figure 4. The numbers printed along the left hand margin refer to the attribute numbers. As you proceed to the right through the tree you

will observe numbers greater than 26. These correspond to the clusterings that takes place from one step to the next. For example, the number 27 shown at the juncture of 25 and 22 means that the first attribute clustered together should be 25 and 22 (this is the most similar pair). This combination is then considered as a new attribute which is later combined with the attribute 30 (itself a combination of 23 and 24) to form the attribute 31. This is later combined with attribute 2 to form attribute 40, etc.

As discussed in Chapter 1 a decision has to be made as to how many clusters (superattributes) are desired. All hierarchical methods will continue clustering until there is a single cluster. In order to decide on the number of clusters (and their composition) one need only image drawing a vertical line through the tree at various places. Each interesection of the tree with the vertical line results in a cluster. For example, the vertical line at the point A results in the 6 clusters shown in Table III.

Table II. Similarity Matrix for Superattribute Determination

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1																										
2	4																									
3	8	1																								
4	7	1	11																							
5																										
6					8																					
7					8	12																				
8		1			3	3	3																			
9					10	10	10	3																		
10			1	1				3																		
11					6	6	6	2	7																	
12					1	3	3	3	1	2		5														
13					1	3	3	3	1	2		5	12													
14					2	5	5	2	4			5	9	9												
15								3		9																
16					3	5	5		4	1	6	3	3	4	3											
17					2	1	1	1	2		5	4	4	3		2										
18					1	1	1	1	1		5	4	4	3		1	9									
19					9	5	5	2	7		3					2	2	1								
20	2	3	1	1	5	1	1	1	3		2						2	1	8							
21					1				1		1						2		3	3						
22	2	8						1												2	2					
23	2	7						1												2	3	11				
24	3	6						1												2	3	10	11			
25	2	1																		2	2	12	11	10		
26					1												1	1	2	2	9	3	4	4	3	

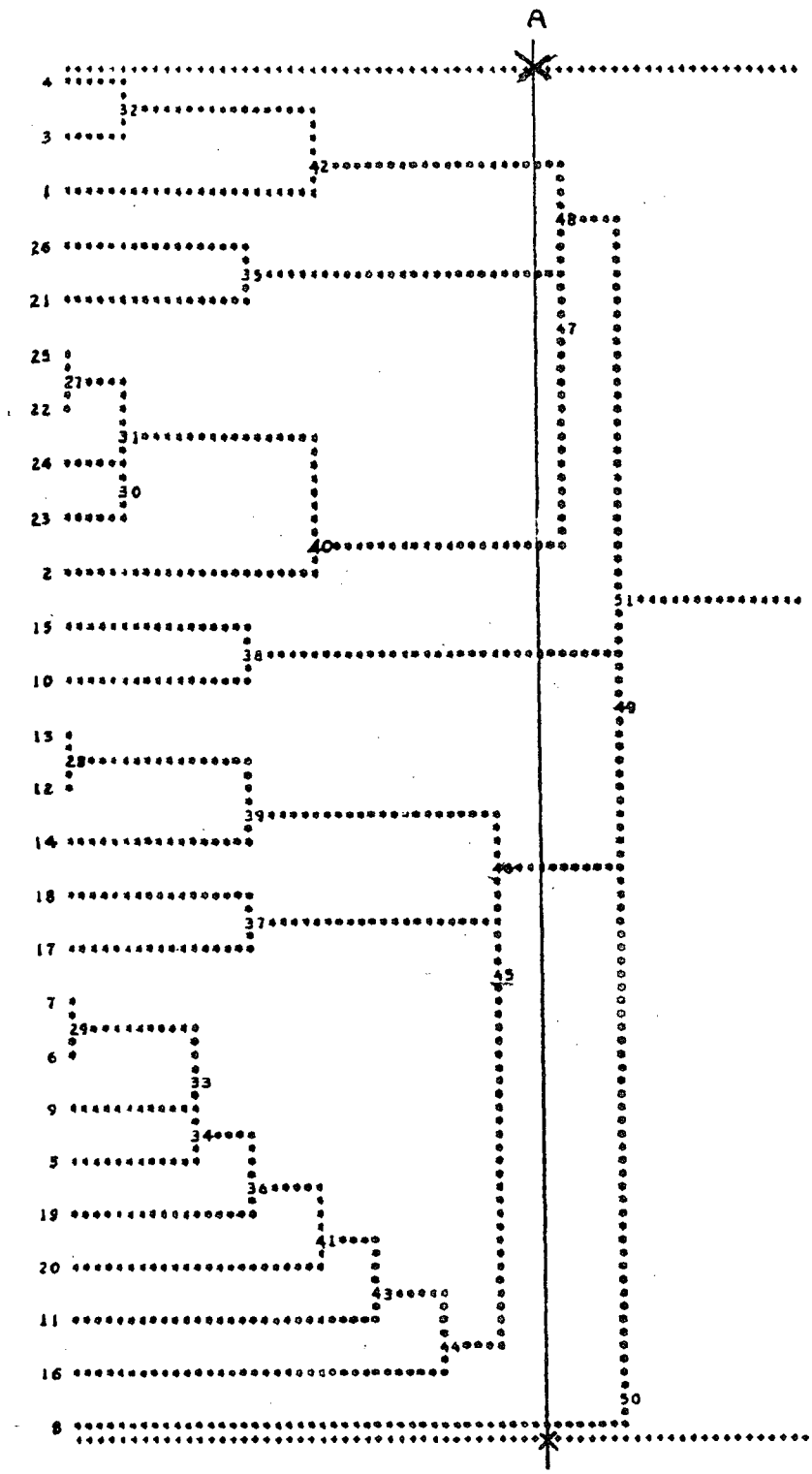


Figure 4. Tree for 26 Attributes

The superattributes used in the utility study are those shown in Table III. A careful examination of the attributes which comapre the clusters shows that the results so obtained are intuitively agreeable. The names supplied to the superattribures are somewhat natural descriptions of the clusters obtained.

Table III. Superattributes

Superattributes	Component Attributes
Facility of movement	1. Portability 3. Time to set up 4. Time to take down
Facility of Use	5. Ease of assigning aircraft to targets 6. Number of aircraft controlled 7. Number of targets 9. Mission flexibility 11. Time to locate and acquire aircraft 16. Aircraft attack throughput 17. Ease of adjustment 18. Accuracy of feedback 19. Ease of operation 20. Man-machine compatibility 12. Accuracy of tracking 13. Accuracy of delivery 14. Range
Survivability	10. ASRT Survivability 15. Aircraft vulnerability
Learning	21. Training requirements 26. Documentation
Readiness	2. Durability 22. Reliability 23. Maintainability 24. Supportability 25. Availability
Communications	8. Communications

3. ANALYSIS OF ARMY TANK DATA

A. DATA STRUCTURE

In order to illustrate the nonhierarchical clustering methodology, principal components analysis, and discriminant analysis data on Army tanks from eight different countries were taken from Jane's Book of Weapon Systems (1979-80). A total of twenty-four tanks were included in the data array with observation on each of 10 variables. The 10 variables are listed below:

(1) Weight (ton)

(2) Length (meter)

(3) Width (meter)

(4) Height (meter)

(5) Road Speed (kilometer per hour)

(6) Trench Crossing (meter)

(7) Ground Pressure (Kg/cm^2)

(8) Maximum Armament (rounds)

(9) Ground Clearance (meter)

(10) Power to Engine Ratio (BHP/ton)

The twenty-four tanks and the associated countries are shown below:

Identification Number	Type/Name	
11	T-62	
12	T-54	U.S.S.R.
13	T-10	
14	ASU-85	
15	MK-5/Chieftain	
16	MK-3/Vickers	
17	MK-13/Centurion	U.K.
18	CVR(T)/Scorpion	
19	XM-1	
20	M60A2	
21	M60	U.S.A.
22	M48	
23	M47	
24	PZ61	
25	PZ68	SWITZERLAND
26	STRV-103	SWEDEN
27	Ikv-91	
28	TYPE61	JAPAN
29	TYPE74	
30	Leopard 2	
31	Leopard	W. GERMANY
32	TAM	
33	AMX 30	FRENCH
34	AMX 13	

We conjecture that a cluster analysis of the tank data will result in clusers corresponding to nationality since the nations may have different emphasis on the variables in the design of their tanks.

B. NONHIERARCHICAL CLUSTER ANALYSIS OF TANK DATA

(1) The MIKCA Algorithm

The specifie algorithm chosen for the nonhierarchical cluster analysis for the tank data is the MIKCA (Multivariate Iterative K-MEANS Clustering Algorithm) program written by Douglas J. McRae as a part of his doctoral dissertation at the University of North Carolina, Chapel Hill.

Reference to the flow chart in Figure 5 will aid the reader in following discussion of the algorithm. Inputs to program are the data matrix, an estimate for g (the number of clusters), and choice of criterion and distance functions.

In the first step, preliminary calculations are made, such as the variable means and standard deviations, as well as the cross product matrix T. The next step forms the initial cluster centers. Then each of the other observations is assigned to the nearest cluster. Euclidean distance is used for this initial phase, and the cluster centroids are recomputed after each observation is assigned to a group. The observations are considered in the same order as they were input. After all of them have been assigned to clusters, the criterion value is computed. This initial cluster-finding technique is referred to as a one-pass K-MEANS procedure.

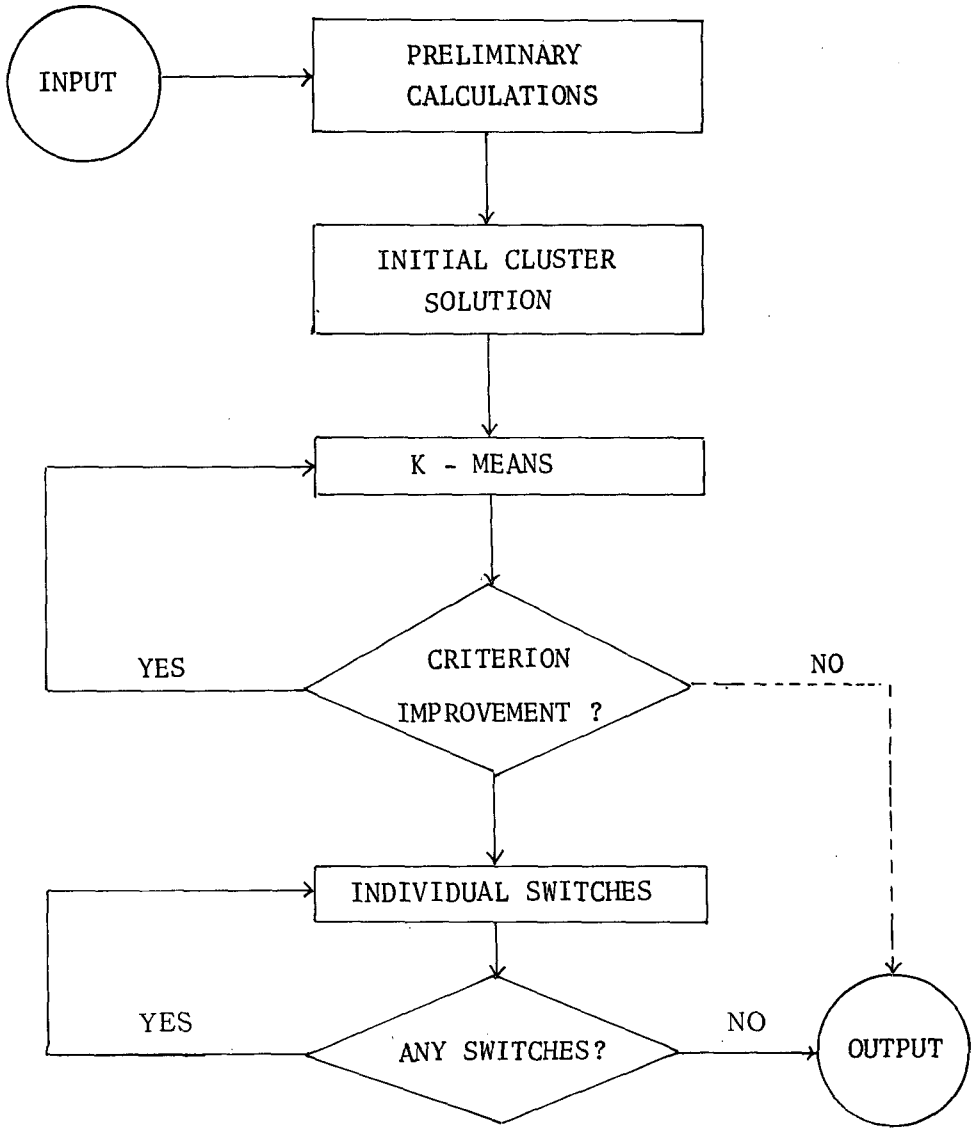


Figure 5. MIKCA Flow Chart

It is performed three times, and the solution which yields the best criterion value is chosen as the initial cluster solution.

After the initial solution has been found, the program advances to the iterative K-MEANS phase where the observations are again considered in the order in which they were input to the program. It is this phase where the user's choice of distance function is used. The distance from each observation to each cluster centroid is again computed, this time with the user's distance function, the assignment to the closest centroid being made and the centroid updated to reflect its new membership. After considering all n observations in this manner, the new criterion value is checked for possible improvement during the K-MEANS iteration. As long as the criterion value improves, the K-MEANS procedure is repeated; if the criterion fails to improve then the MIKCA algorithm goes to the next step, the individual switches section. Note the importance of the order of consideration of the observations. The order is important because the cluster means are recomputed after each observation is reassigned.

In the individual switches phase, consideration is given to moving each observation to every other cluster, the move being made if and only if an improvement in the value of the criterion results. An elaborate labelling procedure provides a unique order in which to consider each observation. This procedure continues until a complete pass through the data is made with no changes in cluster membership.

The MIKCA alogorithm provides the following options for distance and criterion functions.

Ceiterion

(a) Minimum trace W

(b) Minimum determinant W

(c) Maximum largest order of $|B - \lambda W| = 0$

(d) Maximum sum of roots of $|B - \lambda W| = 0$

Distance

(a) Euclidean

(b) Weighted Euclidean

(c) Mahalanobis

A complete computer program is available by author.

(2) Cluster Results for Tank Data

For clustering of the tank data we selected the minimum trace W ceiterion and the weighted Euclidean distance function. The algorithm automatically provides weights for the weighted Euclidean distance function. The results of the clustering with four clusters are shown in Table IV.

The conjecture of clustering by nationalities is supported by the results. The three Soviet tanks make up one cluster and the two British and four of the United States tanks were found to be similar. A third cluster consists of four tanks which are very lightweight. The final cluster consists of the rest of the tanks, including tanks of United States allies from West Germany, France,

Sweden, Switzerland and Japan.

A natural question to ask after observing the results of a cluster analysis is what variables most strongly influence the clustering that was observed. A clue is provided by the composition of the cluster containing all of the lightweight tanks. This suggests that weight is an important distinguishing feature. This is examined in the principal components analysis and the discriminant analysis which are not discussed in this article.

4. CONCLUSION

The multivariate analysis techniques of cluster analysis are useful in real world problems for examining observations on each of several dimension. However, using in combination with principal components analysis and discriminant analysis will bring more analytical results since each of the techniques is related mathematically to the others, and each complements the other in explaining the data.

Computer software is readily available in many sources. The software used in this article for hierarchical clustering was from the IMSL package. For nonhierarchical clustering, I used the FORTRAN program developed by McRae (16). All of this software is readily available and documented at the office of systems analysis in MND.

Table Ⅳ. THE FINAL CLUSTER SOLUTION

	OBSERVATIONS
CLUSTER 1 SIZE = 11	16, 19, 24, 25, 26, 28, 29, 30, 31, 32, 33
CLUSTER 2 SIZE = 6	15, 17, 20, 21, 22, 23
CLUSTER 3 SIZE = 4	14, 18, 17, 34
CLUSTER 4 SIZE = 3	11, 12, 13

BIBLIOGRAPHY

1. Aiken, J. W., "Development of Cluster Analysis for Student Opinion Data," Master's Thesis, Naval Post-graduate School, Monterey, CA, 1979.

2. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1958.

3. Anderberg, M. R., Cluster Analysis for Applications, Academic Press, 1973.

4. Barr, D. R., and Richards, F. R., Utility Assessment Methodology (Report on Relative Utility Score of the AN-TPQ/27). System Exploration Inc., Monterey, CA, 1980

5. Eisenheis, R. A., and Avery, R. B., Discriminant Analysis and Classification Procedures, Lexington Books, 1972.

6. Friedman, H. P., and Rubin, J., "On some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, Vol. 62: 320, Dec. 1967.

7. Giri, N. C., Multivariate Statistical Inference, Academic Press, 1977.

8. Gnanadeskkan, R., Methods for Statistical Data Analysis of Multivariate Observations, Wiley, 1977.

9. Green, P. E., and Carroll, J. D., Mathematical Tools for Applied Multivariate Analysis, Academic Press, 1976.

10. Hartigan, J.A., Clustering Algorithms, Wiley, 1975.

11. Kandell, M. G., The Advanced Theory of Statistics, Vol. III. Charles Griffin and Company, 1947.

12. Keeney, R. L., and Raiffa, H., Decisions with Multiple Objectives, Wiley, 1976.

13. Klecka, W. R., "Discriminant Analysis," SPSS (Statistical Package for the Social Sciences), McGraw Hill, 1975.

14. Kim, J. O., "Factor Analysis," SPSS (Statistical Package for the Social Sciences), McGral Hill, 1975.

15. MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations," Paper presented at Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California, 1965-1966.

16. McRae, D. J., Clustering Multivariate Observations, Doctoral Dissertation, University of North Carolina, Chapel Hill, N.C., 1973.

17. Morrison, D. F., Multivariate Analysis: Technique for Educational and Psychological Research, Wiley, 1971.

18. The IMSL Library, Vol. 3, IMSL, Inc., 1979.