TECHNIQUES FOR MEASURING THE STABILITY OF CLUSTERING: A COMPARATIVE STUDY*

Vijay V. Raghavan
Computer Science Dept.
University of Regina
Regina, Sask., Canada

and

M.Y.L. Ip
Datatron Corp.
Lethbridge, Alta., Canada

ABSTRACT

Among the significant factors in assessing the suitability of a clustering technique to a given application is its stability; that is, how sensitive the algorithm is to perturbations in the input data. A number of techniques that appear to be suitable for measuring the stability of clustering have been published in the literature. The details about each of these measures, such as a description of the steps involved in their computation and an identification of precisely what they measure, are presented. These measures are considered in the context of analysing the stability characteristics of clustering techniques and are compared using a framework developed for this purpose. The question of generalizing some of these measures is addressed and the measures are also analyzed to identify conditions under which they can be reduced to one another.

I INTRODUCTION

In many fields of study such as life sciences, social sciences, and information retrieval, a specialist finds himself confronted with a large number of entities (organisms, persons, documents) each of which is represented by a finite number of attribtues or features. An

* This research has been supported in part by a grant from Natural Sciences and Research Council of Canada.

important step in analyzing and understanding such data may often consist of classifying or clustering this set into "homogeneous" groups.

Many clustering techniques are now available. Some of them are graph theoretic, while the others have variously been described as decision theoretic, enumerative, or corridor and linear adaptive (Salton, 1975). A review of these methods can be found in Bonner (1964), Ball (1965), Johnson (1967), Lance and Williams (1967 a, b), Cormack (1971), Jardine and Sibson (1971), Watanabe (1972), Sneath and Sokal (1973), Yu (1974), Day (1977) and Matula (1977).

There are a number of factors that determine the suitability of a clustering technique to a given application. The most important of them is, of course, the effectiveness of the classification generated in the context of the application. It is also common practice to evaluate a clustering method in terms of other factors such as computational efficiency, whether or not the resulting classification differs depending on the order in which the objects are processed and if the same clusters would be obtained if the scale of certain values is altered. Another significant factor in assessing the suitability of a clustering algorithm is its stability; that is how sensitive the algorithm is to perturbations in the input data. Very few researchers have considered the evaluation of different clustering techniques from the point of view of their stability (Rand, 1971; Yu, 1976; Corneil and Woodward, 1978; Raghavan and Yu, 1981). Consequently, very little has been published about techniques for measuring stability of clustering. On the other hand, a number of studies have considered the problem of how to compare classifications (Sokal and Rohlf, 1962; Farris, 1969; Jackson, 1969; Arabie and Boorman, 1973; Boorman and Olivier, 1973; Rohlf, 1974; Day, 1979). Since the measurement of change in classification, as a consequence of perturbations in the input data, is fundamental to the assessment of stability, the earlier

work on the comparison of classifications is of particular interest. The primary aim of this paper is to consider the various measures referred to above in the context of analysing the stability characteristics of clustering techniques. This goal is accomplished by describing each of the techniques, explaining (when necessary) how they can be used to measure stability of clustering, and providing a comparison of these measures.

The remainder of the paper is organized in the following manner. In section II, a few essential definitions relating to clustering methods are introduced. A diagram which gives readers an overview of the clustering techniques is also presented. In section III we point out the importance of a stable classification from the point of view of document retrieval systems. A detailed description of the techniques found in literature, which can be used to measure stability of clustering, is presented in section IV. These measures are then compared to each other and evaluated in the context of assessing clustering stability in section V. Finally, section VI provides a summary of the findings of this paper.

II BASIC DEFINITIONS

In this paper, the terms clustering, classifying, or clustering technique refer to the process by which the entities, or the attributes characterizing them, are placed into groups such that the objects (attributes or entities) within a group are more strongly related to each other than those in different groups. These homogeneous groups are referred to as classes or clusters. The set of classes or clusters obtained by clustering a given set of objects constitutes a classification. In the placement of objects into clusters, it is generally required that the classification generated be exhaustive of the objects under consideration and that no cluster properly includes any other cluster. If the overlap between any two

clusters in such a classification is null, then it is called a partition.

For the discussions of this paper, it is convenient to place clustering methods into three broad categories. First is the group of clustering methods that have been referred to as graph theoretic. These methods require that the computation of the similarity measure between every pair of objects be the first step in the identification of clusters. This means that an object-object similarity matrix, whose $(i,j)^{th}$ element represents the degree of closeness between the i^{th} and j^{th} object, needs to be created. Then, a threshold is applied and the value of the $(i,j)^{th}$ element is made 1 if the corresponding similarity value is greater than or equal to the threshold; the element is made zero, otherwise. This matrix is referred to as the adjacency matrix. A graph is then associated with the adjacency matrix where each vertex in the graph corresponds to an object, and an edge is associated between two vertices if the element in the adjacency matrix for the corresponding objects is 1. Clusters are then identified by choosing one of various graph theoretic constructs available for specifying the conditions under which a set of vertices would be placed in the same cluster.

The second group of clustering methods considered here are due to Jardine and Sibson (1971). Informally, the clustering schemes analyzed by Jardine and Sibson, consist of first obtaining a "target" similarity matrix from the object-object similarity matrix. Then, from the target similarity matrix, a classification is found by employing the maximal complete subgraph (MCS) clustering method. The formation of the target similarity matrix would, of course, depend on the type of classification desired. Although these methods make use of graph theoretic notions, the overall algorithm is quite distinct from that of the first group.

Some clustering methods (enumerative, decision theoretic, etc.),

however, do not require the computation of the similarity or the adjacency matrices and the classification can be computed more directly from the input. The third and final group, thus, includes all methods not included in the other two.

The categorization described above is summarized in Figure 1. In the figure, the path ABCE represents the operations involved in any graph theoretic clustering technique. The path ABDE corresponds to the class of clustering techniques devised by Jardine and Sibson (1971). All other clustering methods are characterized by the path AE.

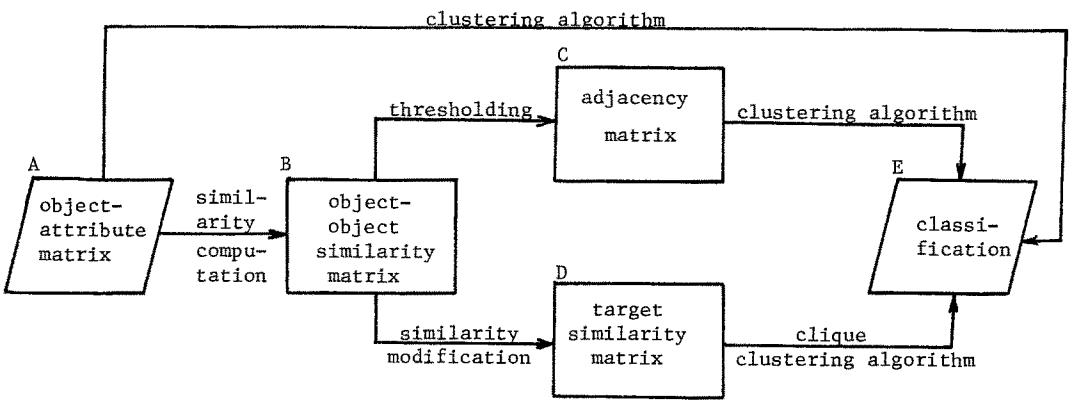


Figure 1. An overview of clustering techniques.
In the figure, boxes A and E represent the input and the output, respectively, of the clustering process. Each arrow indicates an operation, whereas the boxes at the tail end and at the head correspond respectively to the input and output of the operation.

III MOTIVATION FOR THE CURRENT WORK

It was mentioned that there are a number of factors that determine the suitability of a clustering technique to a given

application and that stability is one among them. The concern about stability stems from the fact that perturbations to the input do occur, as a result of errors made in the parametric representation (choice of attributes and their values) of objects and in the conversion of this data into machine readable form. In this section, in order to demonstrate that the study of stability is important, we first look at the kinds of errors that can arise in the context of information retrieval systems*. Secondly, we consider the issue of measuring stability and difficulties thereof. Finally, the objectives of the current work are reviewed in the light of the discussions referred to above.

III. 1. The sources of input errors

The following kinds of errors can arise in the initial processing and preparation of document collections:

(a) In handling large document collections, the information from which their descriptions are obtained would have to be transcribed from their source and, then, be converted into a form suitable for processing by a computer. During this stage of input preparation, various kinds of clerical errors may occur.

(b) A descriptor which is not important or does not appear in a document may be mistakenly assigned.

(c) The set of descriptors that properly represent the content of a document may not be unique. That is, although two indexers (or, indexing strategies) analysing the same document may produce descriptions which agree with each other substantially, it is nevertheless likely that there will be points of disagreement.

(d) In a dynamic environment in which additions and deletions of

* In these systems, the database is a collection of documents or texts. Both documents and queries are represented by a set of descriptors. In response to a query, the closeness of the various documents to the queries are determined using some kind of a best match criterion and a list of references are provided.

documents occur quite frequently, these changes to the collection can be viewed as a kind of perturbation. The dynamic aspect would also have implications for the extent to which document descriptions remain valid over time. As new jargon is introduced into a field, earlier descriptions of documents can become incomplete or inaccurate.

It is easily seen that these kinds of perturbations are unavoidable. But, if the clustering method is able to obtain nearly the same classification in spite of these distortions, it would clearly be an asset. In the light of the discussion above, we informally refer to a clustering method as stable if small changes in the input data lead only to small changes in the classification generated.

III. 2. Difficulties in the evaluation of stability

There are essentially two aspects that a measure of stability must deal with. One aspect is the evaluation of changes in the input data due to errors, and the other is the measurement of the differences in the resulting classifications.

Assuming that a collection of documents is viewed as an array, changes to a document-term array can be measured quite easily. Suppose that two document arrays are given, representing differing descriptions of the same document collection. There is no difficulty in identifying a particular document in one array with the corresponding document in the other array. Thus, over the whole collection, a measure of difference between the two arrays may be obtained by first computing the corresponding object-object similarity matrices and, then, determining the similarity between these two matrices by using a measure such as Kendall's (1938) coefficient of agreement.

There are, however, difficulties in using a similar approach for comparing classifications, even though a classification, like the

document collection, can be represented by means of a binary array. (In a classification array, C, the element C_{ij} indicates the presence or absence of object j in class i.) Suppose that C and C′ are two classification arrays produced from two document arrays, D and D′, which are differing representations of the same document collection. Since names or labels given to each class in C and C′ are entirely arbitrary, it is not possible (as could be done for document arrays) to identify each class of C as corresponding to a specific class in C′. This means, more ingenious approaches are required in order to obtain an object-object similarity matrix on the basis of a classification array.

Thus, it is seen that in an information retrieval environment various kind of errors can occur and that the need for identifying stable clustering methods is real. In order to perform such evaluations one must first have techniques by which stability can be measured. However, comparison of classifications - a process which is inherent to any measurement of stability - is not something that is easily done. In view of these facts, there exists a need to take stock of what is known about evaluating clustering stability and to develop a proper framework in which the work in this area can be well understood. In this paper we attempt to fulfill the above need.

IV EARLIER WORK

Before the techniques for measuring stability of clustering are compared, the earlier work in evaluating stability and the related area of comparison of classifications is reviewed.

IV. 1. Comparison of Classifications

One of the first suggestions put forward for the comparison of different classifications was by Sokal and Rohlf (1962). They propose

a method for comparing hierarchical classifications* which are represented by dendrograms (diagrams of relationships). The purpose of a dendrogram is to show the level at which two or more objects combine to form a common cluster. To illustrate, let us consider 5 objects whose object-object similarity matrix is as given below:

O_2	0.6			
O_3	0.4	0.8		
O_4	0.1	0.5	0.7	
O_5	0.1	0.2	0.2	0.3
	O_1	O_2	O_3	O_4

Suppose that the clusters corresponding to a given threshold are defined (borrowing a graph theoretic terminology) as the connected components (CC's) of the associated graph. Then, the dendrogram for this situation is as shown in Figure 2. In a dendrogram the abscissa has no particular meaning. The ordinate, on the other hand, represents similarity values. In the example given, O_2 and O_3 join at level 0.8, O_4 combines with O_2 and O_3 at level 0.7, O_1 combines with O_2, O_3 and O_4 at level 0.6 and, finally, all the objects form a single cluster at level 0.3.

A summary representation of the level at which the various pairs of objects join is obtained as explained below and dendrograms can be compared on the basis of such a representation.

The range of similarity values along the vertical axis is divided into a suitable number of equal intervals. Suppose that the number of intervals is N and the similarity values are in the range between 0 and 1. Let the code number of the (similarity value) interval ((i-1)/N,i/N) be i, for $1 \leq i \leq N$. Then, the cophenetic value of a given

* In a hierarchical classification there are many levels of classifications, and the clusters in a given level are cohesive than those of any higher level. That is, any cluster at a given level is a subset of some cluster at each higher level.

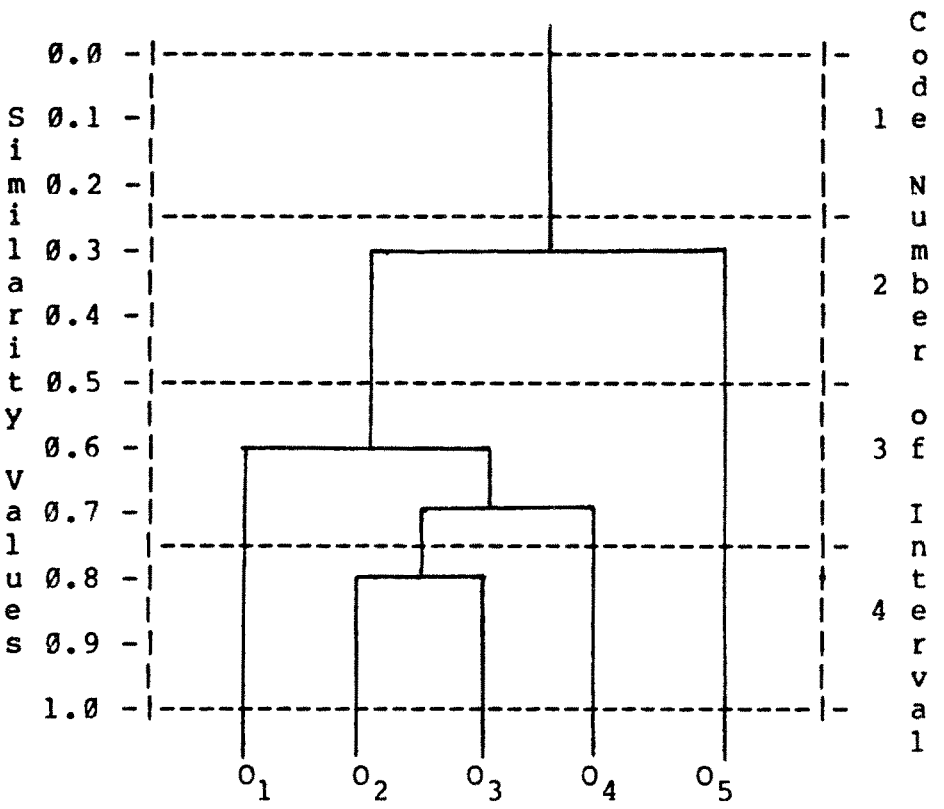


FIGURE 2. A dendrogram to illustrate the computation of cophenetic values.

pair of objects is defined to be the code number of the interval that contains the similarity value at which the objects join in the dendrogram. Using this scheme, a matrix containing the cophenetic value for every pair of objects can be generated. The values generated accordingly are given by the elements below the main diagonal of the following matrix. Note that this is an object-object similarity matrix which is derived on the basis of how the objects have been clustered.

O_1	X	4	4	3	3
O_2	3	X	2	3	5
O_3	3	4	X	3	5
O_4	3	3	3	X	4
O_5	2	2	2	2	X
	O_1	O_2	O_3	O_4	O_5

Thus, given two classifications, the corresponding (similarity) matrices of cophenetic values are generated and a measure of agreement (or disagreement) between them is determined by calculating the product moment correlation coefficient, which measures the extent to which there is a linear relationship between the two sets of cophenetic values.

Farris (1969) has suggested an alternative to the cophenetic value, which is called the cladistic difference. Consider a tree diagram corresponding to a dendrogram in which the external nodes are the objects, and each internal nodes represents the merging of two or more lower level clusters. Then, the cladistic difference between two objects is the number of edges in the path between them on the tree diagram. These values, for our example, are given by the elements above the main diagonal of the matrix shown above.

In Jackson's (1969) work, a measure which reflects the extent to which a classification truly represents the data from which it is

derived is developed. Again, let D denote the object-attribute binary array and C the object-class binary array which is obtained from D by employing some clustering method. Let S and T denote the object-object similarity matrices obtained respectively from D and C by applying a similarity function to all the object descriptions considered pair-wise. Then, the assessment of the discrepancy between S and T is made by checking whether on not sign(S(i,j) - S(k,l)) = sign(T(i,j) - T(k,l)), for each distinct combination of i, j, k and l that represent four different objects. In other words, whether or not a pair of objects which are more similar, by virtue of their initial attributes, than another pair of objects, are also more similar as indicated by the clusters in which they have been jointly placed. Thus, the greater the number of cases in which this condition holds, the better is a classification. The discrepency measures proposed by Jackson, which combine the numerous checks mentioned above to a single value, except for minor differences, can be thought of as being the complement of Kendall's (1938) coefficient of agreement.

Borko et al. (1968) suggested constructing simple contingency tables for comparing non-hierarchic, non-overlapping classifications. An element, f_{ij}, in the table gives the number of objects in class j in the first classification that are in class i in the second classification, for some arbitrary labelling of the clusters in the two classifications. For example, if Y and Y' are two classifications of six objects {a,b,c,d,e,f} given by

$$Y = \{ \{a,b,c\},\{d,e,f\} \} \quad \text{and}$$

$$Y' = \{ \{a,b\},\{c,d,e\},\{f\} \} ,$$

then the contingency table will be

		Y classes	
	Y'	2	0
classes		1	2
		0	1

Using the table, the degree to which one can predict the class in which an object will fall in one classification, knowing only its class in the other classification can be measured by performing various contingency coefficients and tests of independence (e.g. using, χ^2).

Since a classification can be a partition, methods proposed for comparing differenct partitions of the same set of objects are of interest. The metric, which possess intuitively desirable properties, has been proposed as a model of distances between partitions (Arabie and Boorman, 1973). Day (1979) has studied metrics on partitions comprehensively. By using models for methodical enumeration of metrics, he identifies twelve metrics. Two of these are classified as pair bond (PB) metrics, while the rest fall in the category of minimum-length sequence (MLS) metrics. Four of the twelve metrics, it is suspected, are difficult to compute, but efficient algorithms for the remaining eight metrics exist and exhibit time complexities ranging from $O(n)$ to $O(n^3)$, where n is the number of objects in the partitions. Boorman and Olivier (1973) have shown that partition metrics can be used to construct metrics on various types tree-like classifications. Thus, the metrics mentioned above are also relevant to the comparison of hierarchical classifications.

IV. 2. Measurement of Stability

Rand (1971) has proposed a method to measure the similarity between two different classifications (actually, partitions) of the same set of objects. The measure essentially considers how each pair of objects is assigned to clusters in the two classifications. If a pair of objects is placed together in a (some) cluster in each of the two classifications, or if the objects in the pair are assigned to different clusters in both classifications, then such pair of objects is said to be "similarly placed". In contrast, an object pair is

defined to be "differently placed" if the pair is in the same cluster in one classification and the objects in the pair are in different clusters in the other. A measure of similarity between two clusterings, Y and Y′, can be defined as c(Y, Y′), and is equal to the number of pairs of objects which are similarly placed normalized by the total number of object-pairs.

The following example illustrates the calculation of c between two classifications, Y and Y′, of six objects. Let Y = {(a,b,c), (d,e,f)} and Y′ = {(a,b), (c,d,e), (f)}, then the object-pairs are tabulated as follows:

object-pair	ab	ac	ad	ae	af	bc	bd	be	bf	cd	ce	cf	de	df	ef	TOTAL
similarly placed	*		*	*	*		*	*	*				*	*		9
differently placed		*				*				*	*				*	6

A total of nine pairs being similarly placed out of a possible 15 gives c(Y, Y′) = 0.6.

It is clear that the measure of similarity c, ranges from 0, when the two classifications have no similarities at all, to 1, when they are identical.

Rand suggests, in his paper, that the measure c can be used to study various characteristics that researchers would like to investigate, prior to choosing a particular clustering method. One such characteristic mentioned is the sensitivity of a method to perturbation of the data. Corneil and Woodward (1978) choose stability as one of the properties in their comparison of three clustering methods. In this case, Rand's measure is used for measuring stability. Thus, if Y and Y′ correspond to classification obtained, respectively, for the unperturbed and perturbed data, then the larger the value for c, the more stable is the clustering method in question.

In a recent study Yu (1976) proposes a method for measuring the amount of disturbance in classification due to small changes in the input data. The measure is developed in particular reference to graph theoretical clustering strategies. The proposed measure estimates the amount of change in a set of
clusters by the minimum number of 'operations' required to restore the set of modified clusters to the original ones. Implicit in the above statement is the assumption that the change in the data is so small that there is a 1-1 correspondence between the vertices of the modified graph and those of the original graph. In other words, neither any of the original objects is lost nor any new ones created. Therefore, the operations consist only of the addition and/or deletion of edges needed to restore the set of modified clusters to the original ones.

More precisely, let GP= (V,E) be the graph that would represent the object-object similarities if there had been no errors in the input data. Let G* = (V,E*) denote the graph actually obtained as the result of some perturbations in the input. That is, E* is obtained by deleting some edges from E and adding some edges from E to E. Thus, edge deletions come from the original graph, whereas edge additions are from the complement graph. Given some method of defining a classification, suppose G** = (V,E**) denotes a graph which is obtained through minimum number of changes to G* such that G and G** have an identical set of clusters, then, the amount of change is specified by the expression $|(E^{**}-E^{*}) \bigcup (E^{*}-E^{**})|^{+}$. This concept is illustrated below using Figure 3.

The changes to G due to errors are the addition of edge $(0_3, 0_7)$ and the deletion of $(0_1, 0_2)$ and $(0_4, 0_5)$. Let the clusters be defined as the CC's. Clearly, the removal of $(0_3, 0_7)$ and the addition of one of the edges $\{(0_1, 0_2), (0_1, 0_5), (0_4, 0_2), (0_4, 0_5)\}$

+ If A is a set of p elements, the $|A| = p$.

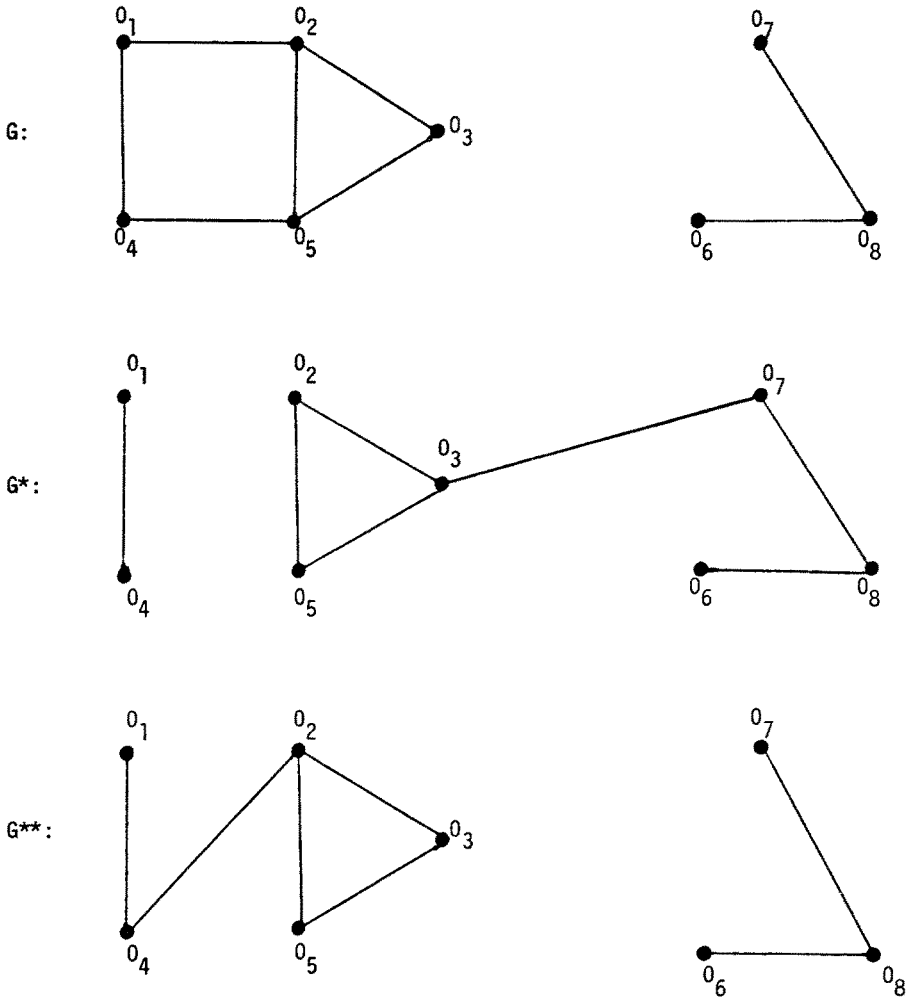


FIGURE 3. An initial, a perturbed, and a restored graph to illustrate the amount of work.

are sufficient to restore the clusters (refer to G**), and the work needed for restoration is 2. Note that, certain edge changes (adding $(0_3, 0_7)$ or deleting $(0_6, 0_8)$) would affect the resulting clusters and yet certain other changes (say, deletion of $(0_2, 0_3)$) do not alter the clusters obtained. In this sense, the measure takes the structure of the graph into account.

Yu finds that clusters defined as the MCS's require the maximum number of operations and hence the least stable. In fact, in this case, the corrected graph must be identical to the initial graph. He compares the effect of simple matching and cosine similarity functions experimentally and concludes that for both cluster defining methods tested (CC and MCS), clusters produced on the basis of cosine function are less stable than those obtained using simple matching.

Raghavan and Yu (1981) use the measure proposed by Yu to compare the stability characteristics of a number of families of graph theoretic clustering schemes. The connected component method is shown to be the most stable of all graph theoretic clustering methods that possess a certain property, and the maximal complete subgraphs method is found to represent the worst possible case in terms of stability. Furthermore, it is shown that certain families of graph theoretic clustering algorithm are such that as one proceeds from the method producing the most narrow clusters (MCS) to those producing relatively broader clusters, the clustering process remains at least as stable as any method in the previous stages.

V COMPARATIVE EVALUATION OF TECHNIQUES FOR MEASURING STABILITY

Of the various techniques reviewed in the previous section, those considered under the heading of Comparison of Classifications are of potential interest to the assessment of clustering stability. Since they have been proposed in a broader context, first we consider how

these measures might be used for measuring cluster stability. Secondly, we analyse the process in reference to Figure 1, and summarize the differences in the techniques in terms of the approach employed for measuring stability. This analysis also provides some possibilities for other approaches that might be used. Finally, some special features and limitations inherent to the measures considered are outlined. The discussion in this final subsection, then, lead to some improvements and generalizations.

V. 1. Measuring stability using the techniques for comparing classifications

The techniques for the comparison of classifications presented in section IV.1 are those by Sokal and Rohlf (1962), Jackson (1969), Day (1979) and Borko et al. (1968). Generally speaking, all these measures can be used for assessing clustering stability.

These methods would likely be used in an experimental setting, where the amount of error associated with the input (object-attribute matrix) would be introduced in a controlled fashion. Two classifications would be generated, one for the correct input and the other for the perturbed input. The classifications can then be compared using one of the above measures and the effect of input errors on classification can be determined.

Thus, the adoption of methods studied by Day or that proposed by Borko is straight forward since these measures directly deal with the similarities and differences between the classifications given. In the case of the method by Sokal and Rohlf, first the matrix of cophenetic values, which is essentially an object-object similarity matrix, would be derived from each of the resulting classifications. Then, the two matrices are studied to determine the extent to which the ranking of the similarities between the various pairs of objects in one matrix coincides with the corresponding ranking in the other

matrix.

The use of Jackson's method for measuring stability of clustering is similar to that of Sokal and Rohlf in the sense that object-object similarities based on the two resulting classifications is first obtained. At this point, one approach that might be used to determine the stability of the clustering method used is to compare the object-object similarity matrices. Alternatively, the use Jackson's method can be fashioned more closely after how he proposed to compare classifications.

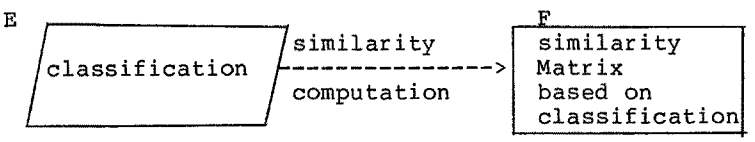
As indicated earlier, Jackson's proposal was to compare the similarity values obtained on the basis of the input object-attribute matrix to those obtained on the basis of the resulting classification. The idea is that this comparison would indicate how well the classification still (after classification) retains the original relationships. This approach suggests that a better classification method is more able to retain the relationships that originally existed. In keeping with this thinking, we might assert that the clustering method is stable if the classification obtained using the perturbed input reflects the relationships in the correct input as well,as does the classification that is obtained on the basis of the correct input. Accordingly, the object-object similarities based on the two classifications would be compared separately with the similarity values obtained from the input object-attribute matrix.

V. 2. A framework for understanding approaches to measuring stability

A number of methods which are available for the evaluation of stability of clustering techniques have been described. In this section, these methods are classified into a number of categories in terms of the approach they employ to measure stability. For this analysis, we characterize the approach of the various methods

considered in reference to Figure 1 of section II.

In the previous section, it was pointed out that in some methods of comparing classifications, an object-object similarity matrix is derived from each of the classifications and, then, the comparison is made in terms of these similarity matrices. In order to correctly describe this case in terms of Figure 1, the following modification to the figure is suggested. A box labeled F is added to the right of box E. An arrow is added which points from box E to box F as shown below:



The arrow stands for the process of computing object-object similarities from classification resulting from the clustering process.

The approaches for measuring stability employed by the various methods can now be summarized.

Approaches to measuring stability:

(i) Compare F to F′

(ii) Compare B, in turn, to F and F′

(iii) Compare E to E′

(iv) Compare C to C′

The labels used with an apostrophe refer to the same entity, except that they are obtained after the input data has been perturbed to reflect the effect of errors introduced. It is easy to see that methods of Sokal and Rohlf, and Farris are of type (i). The use of Jackson′s method in the content of measuring stability would correspond to (ii). The method proposed by Borko, as well as the partition metrics studied by Day, fall into the third category. It turns out, the method proposed by Rand is one of the metrics considered by Day. This has been referred to as the D metric in Day′s

work and elsewhere in the literature (Arabie and Boorman, 1973). Finally, we note that the method suggested by Yu is of type (iv). The measure in this last case, however, depends not only on C and C', but also on the specific graph theoretic construct used to specify clusters.

We conclude this section by pointing out some of other approaches that might also be adopted. These approaches are motivated by the relationship that exists between the co-phenetic value matrix of Sokal and Rohlf and the target similarity matrix (box D) identified in Figure 1. Let us consider again the original object-object similarity matrix from which the dendrogram of Figure 2 is derived and, for that matrix, show the target similarity matrix that would be obtained by Jardine and Sibson's (1971) clustering scheme. It is assumed that the clustering scheme chosen is the one that would lead to the same dendrogram as Figure 2. The original and the target similarity values are shown in the matrix below.

O_1	X	0.6	0.6	0.6	0.3
O_2	0.6	X	0.8	0.7	0.3
O_3	0.4	0.8	X	0.7	0.3
O_4	0.1	0.5	0.7	X	0.3
O_5	0.1	0.2	0.2	0.3	X
	O_1	O_2	O_3	O_4	O_5

The values below the main diagonal are the original object-object similarities and those above the main diagonal are the target similarity values. To illustrate the process by which target similarities are obtained, we explain how the target similarity of (O_2, O_4) changes from 0.5 to 0.7. By the original similarities, the objects (O_2, O_3) join in a cluster at the level of 0.8. Then, O_4 joins this cluster at the level of 0.7, by virtue of the fact that the original similarity between (O_4, O_3) is 0.7. This step results in O_4

also joining with O_2. Thus, even though the original similarity between (O_4, O_2) is 0.5, O_4 is considered to join the cluster having O_2 at level 0.7. In this way, the other changes can be explained.

Now we notice that a striking similarity exists between the target similarity values and the cophenetic values. The process by which one obtains the cophenetic values can be seen simply as a generalization of the thresholding process (see Fig. 1) by which an adjacency matrix is obtained from an object-object similarity matrix. Let this process be referred to as multi-valued thresholding, where the similarity values are broken down to a number of intervals and each interval is mapped to a code number. Thus, in our example, similarity interval (0.25-0.5), (0.5-0.75) and (0.75-1.00) get mapped, respectively, to code numbers 2, 3 and 4. This mapping applied to the target similarity values above yields the matrix of co-phenetic values presented in section IV. 1.

Since, when each unique value in target similarity matrix falls in a different interval, the matrix of cophenetic values is essentially identical to target similarity matrix, it is reasonable to assert that the target similarity matrices can provide a basis for comparison. Consequently, we have the following further possibilities for evaluating stability:

(v) Compare D and D′

(vi) Compare B, in turn, to D and D′

(vii) Apply the thresholding operation to D and D′ and compare the binary (adjacency) matrices that result.

It is also interesting to note that, in the sense of approach (vii), comparing single-level classifications can be viewed as simply a special case of comparing hierarchical classifications.

Several of the methods considered can be varied by using a different function for the comparison of similarity matrices. In this respect, in section III. 2, the possibility of using Kendall′s

coefficient of agreement is suggested, and Sokal and Rohlf (section IV. 1) recommend the use product moment correlation coefficient. Jackson proposed measures denoted as g(-) and g(+), which are closely related to Kendall's coefficient of agreement. Many other choices are possible. The reader is referred to a table presented by Rohlf (1974) in which a number of such coefficients are listed.

V. 3. Comparisons and generalizations

The methods considered in this study can be compared on many respects. In presenting such comparisons, the method for measuring stability of clustering are treated roughly in the order in which they were reviewed.

Sokal and Rohlf's method was proposed specifically for comparing hierarchical classifications. It is well suited to the clustering techniques such as those proposed by Jardine and Sibson. Since one of the intermediate results of these clustering schemes is the target similarity matrix, the computational effort involved in measuring stability would be of $O(n^2)$, where n is the number of objects, for both constructing the matrix of co-phenetic values and determining the product moment correlation coefficient. This method can be used regardless of whether the clustering procedure yields overlapping or non-overlapping clusters in the various levels. The cladistic differences proposed by Farris (1969) could be used, for hierarchical classifications, instead of cophenetic values. But since cladistic differences ignore the relative levels at which braching takes place, the use of this distance measure may not be appropriate for many applications (Rohlf, 1974).

The method proposed by Jackson (1969) is suited to classifications that are commonly encountered in information retrieval applications. That is, classifications which just have a single level and where the classes in the classifications are allowed to overlap.

Jackson also assumes that the input object-attribute matrix is binary. As mentioned earlier, the entity corresponding to box E of figure 1 for this case is a classification array. The object-object similarities, represented by F, are obtained by applying a similarity function (e.g. Tanimoto function) to the object descriptions in E. Again, in reference to figure 1, the correlation of B to F (or B to F′) is more complex computationally since Jackson proposes to compute concordant and disconcordant pairs, not just once for the complete matrices B and F, but at many stages with each stage corresponding to an application of the thresholding operation to these matrices. Thus, while the computational effort depends on the number of objects, it is dominated by the number of distinct values that appear in B. The process can therefore be speeded up by mapping the similarity values in B to another, more suitable scale. The details of the computational procedure can be found in Jackson (1969).

Jackson's method is not really suitable for comparing partitions. The problem here is that, for partitions, the classification array has just a single 1 in the column corresponding to each object. Thus, the object-object matrix derived from this array would also be binary. It is easy to see that correlating such a binary matrix with the object-object similarities derived from the input data can lead to meaningless results. This method can, however, be generalized to be applicable to hierarchical classifications. Rohlf (1974) illustrates how a classification array can be obtained, given a hierarchical classification. Using that approach, the classification array for the dendrogram of Figure 2 would be as follows:

	O_1	O_2	O_3	O_4	O_5
C_1	0	1	1	0	0
C_2	0	1	1	1	0
C_3	1	1	1	1	0

The matrix shows the classes at the lowest level, as well as the clusters that they expand to, as we consider higher levels. The level which has all the objects in a single cluster is not represented. Having constructed the classification array in this way, the method would proceed as before. We find, then, that Jackson's method is also applicable to Jardine and Sibson type clustering schemes, regardless of whether the resulting hierarchical classifications have overlapping or non-overlapping clusters in each level.

The measures studied by Day are partition metrics, and as the name implies they are designed for comparing single-level classifications without cluster overlap. The computational complexity of these metrics and the possibility of their generalization to hierarchical classifications were alluded to in section IV. 1. Among these metrics, the one referred to as the D-metric is of particular interest. This D-metric is the same as that proposed by Rand. It has been shown that the computation of this metric requires time of O(n) for n objects. As we shall see later in this section, this measure also generalizes to the case of single level classifications in which clusters are allowed to overlap.

The stability measure proposed by Yu is designed for the analysis of graph theoretic clustering methods. Although its use is limited to this context, it has the distinct advantage (over the other methods) that the stability characteristics of graph theoretic methods could be studied analytically. In Raghavan and Yu (1981), several such results have been proved. The computational details are not really relevant since this measure is not intended to be used in an experimental setting. This measure of stability is applicable to either overlapping or non-overlapping clusters.

The methods proposed by Rand, and by Sokal and Rohlf can be reduced to Yu's measure in the following way. Referring to Figure 1 again, if we are given E and E', we could construct adjacency

matrices, C and C′, such that for any two objects in the same cluster in E the entry in C is 1, and 0 otherwise; C′ is defined similarly. If Yu's measure is applied to these matrices, assuming clusters are defined as MCS's, the result will be same as that for Rand. In other words, if each of the clusters in E and E′ are imagined to be MCS's of some graph, then Rand's measure is equivalent to computing the number of edge changes required to make the two graphs identical. This correspondence is useful since it suggests a natural way in which Rand's measure can be generalized to measure the distance between classifications that are not also partitions.

Let us now relate Sokal and Rohlf's method to the above. It was mentioned that, when the clustering schemes of Jardine and Sibson are used, the approach can be considered to be one of applying multi-level thresholding to the target similarity matrices, D and D′, and comparing the resulting matrices. It turns out, that if the cluster scheme chosen is equivalent to that of MCS clusters, then the target similarity matrix is same as the original object-object similarity matrix derived from input. Thus, under the special case that, after thresholding the target similarity matrix, the resulting matrix of cophenetic values is binary, the entities that are compared would be C and C′, which is the case for Yu as well. It should be pointed out, though, that the product moment correlation coefficient would not yield the same result as that of Yu. This correspondence, however, may be useful in deciding if some correlation other than the product moment correlation coefficient would be more appropriate.

VI CONCLUSION

The interest in this work stemmed from a desire to study various clustering techniques from the point of view of their stability. This task was rather difficult since many possibilities existed for how clustering stability might be measured and very little was known about

what factors determined which choice to make. This paper alleviates the above problem by analysing and comparing a number of measures available in the literature for evaluating clustering stability. A framework which facilitates the classification of these measures into a number of generic approaches is introduced. Possible generalizations of some of these measures, so that they apply to situations to which they could not initially be used, are also presented.

REFERENCES

Arabie, P. and Boorman, S.A. (1973). Multidimensional scaling of measures of distance between partitions. J. Math. Psych., 10, 148-203.

Ball, G.H. (1965). Data analysis in social sciences: What about details. Proceedings AFIPS. FJCC, Macmillian, New York, N.Y.,533-559.

Bonner, R.E. (1964). On some clustering techniques. IBM J. of Research and Development, 8, 22-32.

Boorman, S.A. and Olivier, D.C. (1973). Metrics on spaces of finite trees, J. Math. Psych., 10, 26-59.

Borko, H., Blakenship, D.A. and Burket, R.C. (1968). On-line information retrieval using associative indexing. Technical Report, Systems Development Corporation, RACD-TR-68-100.

Cormack, R.M. (1971). A review of classification. J. Royal Statistical Society-Series A, 134, 321-367.

Corneil D.G. and Woodward, M.E. (1978). A comparison and evaluation of graph theoretical clustering techniques. INFOR, 16, 74-89.

Day, W.H.E. (1977). Validity of clusters formed by graph-theoretic cluster methods. Mathematical Biosciences, 36, 229-317.

Day, W.H.E. (1979). The complexity of computing metric distances between partitions. Technical Report No. 7901, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.

Farris, J.S. (1969). A successive approximation approach to character weighting. Syst. Zool., 18, 374-385.

Jackson, D.M. (1969). Comparison of classifications. In: Cole (Ed.), Numerical Taxonomy, pp. 91-111, Academic Press Inc., New York, N.Y.

Jardine, N. and Sibson, R. (1971). Mathematical Taxonomy. John Wiley & Sons, Inc., New York, N.Y.

Johnson S.C. (1967). Hierarchical clustering schemes. Psychometrika, 12, 241-254.

Kendall, M.G. (1938). A new measure of rank correlation. Biometrika, 30, 81-93.

Lance, G.N. and Williams, W.T. (1967a). A general theory of classificatory sorting strategies. I. Hierarchical system. computer J., 9, 373-382.

Lance, G.N. and Williams, W.T. (1967b). A general theory of classificatory sorting strategies. II. Clustering systems. Computer J., 10, 271-277.

Matula, D.W. (1977). Graph theoretic techniques for cluster analysis algorithms. In: Van Ryzin (Ed.), Advance Seminar on Classification and Clustering, pp. 95-129, Academic Press Inc., New York, N.Y.

Raghavan V.V. and Yu, C.T. (1981). A comparison of the stability characteristics of some graph theoretic clustering methods. IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-3, 393-402.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. J. of the American Statistical Association, Vol. 66, 846-850.

Rohlf, F.J. (1974). Methods of comparing classifications. Annu. Rev. Ecol. Syst., 5, 101-113.

Salton, G. (1975). Dynamic Information and Library Processing. Prentice-Hall, Englewood Cliffs, N.J.

Sneath, P.H.A. and Sokal, R.R. (1973). Numerical Taxonomy. Freeman, San Francisco, Ca.

Sokal, R.R. and Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. Taxon, 11, 33-40.

Watanabe, S. (1972). A unified view of clustering algorithms. In: Information Processing 71, North Holland Publishing Co., Amsterdam, 149-154.

Yu, C.T. (1974). A clustering algorithm based on user queries. J. of the American Society for Information Science, 25, 218-226.

Yu, C.T. (1976). The stability of two common matching functions in classification with respect to a proposed measure. J. of the American society for Information Science, 27, 248-255.