© Journal of the American Statistical Association
December 1971, Volume 66, Number 336
Theory and Methods Section

Objective Criteria for the Evaluation of Clustering Methods

WILLIAM M. RAND*

Many intuitively appealing methods have been suggested for clustering data; however, interpretation of their results has been hindered by the lack of objective criteria. This article proposes several criteria which isolate specific aspects of the performance of a method, such as its retrieval of inherent structure, its sensitivity to resampling and the stability of its results in the light of new data. These criteria depend on a measure of similarity between two different clusterings of the same set of data; the measure essentially considers how each pair of data points is assigned in each clustering.

1. INTRODUCTION AND STATEMENT OF PROBLEM

Cluster analysis has come to be used as a generic term for techniques which are concerned with the problem: given a number of objects, how to select those which are closer to each other than they are to the rest of the objects. While the general problems of clustering have long intrigued investigators, the necessity and recent availability of large computing power has shaped much of the recent research (see [1, 6, 7], which contain computer-oriented reviews of recent effort). Although some theoretical investigations have been made [e.g., 2, 4, 5], most effort has been directed toward the investigation and development of specific methods in specific situations [e.g., 3, 4, 8, 9]. Much of this research has been directed toward finding natural definitions of the term closer and developing and evaluating clustering methods in these terms. This article assumes that every definition of "closer" is natural for some situation and therefore that the problem can be considered without this aspect. This article focuses instead on the development of procedures for evaluating clustering methods in objective terms of how they cluster.

The specific clustering problem discussed is defined as the study of the triplet (X, Y, m). In this notation X represents the set of N objects (or points) to be clustered, $X=\{X_1, X_2, \cdots, X_N\}$, and Y, a specific partitioning of these objects into K disjoint sets. This partitioning will be called a clustering and written as a set of clusters, $Y=\{Y_1, Y_2, \cdots, Y_K\}$, where each cluster is a set of the given points, $Y_k=\{X_{k_1}, X_{k_2}, \cdots, X_{k_{n_k}}\}$, with $\sum n_k=N$ and $n_k \geq 1$ for $k=1, 2, \cdots, K$. The set of all such partitions, for all $K=1, 2 \cdots, N$, of a given set of N points will be denoted by $\mathcal{Y}$.

The symbol m is used for the method of choosing a particular Y given the set X. In general, clustering methods have two components, a criterion and a technique. The criterion assigns to each clustering a numerical value which indicates its relative desirability in the context of the given method. The technique selects a particular subset of $\mathcal{Y}$ to be searched for a clustering which minimizes (maximizes) the given criterion. The technique is an essential part of any operational method, since even for moderate N, the number of elements of $\mathcal{Y}$ precludes examination of all possible clusterings.

2. A STATISTIC FOR INVESTIGATION OF THE CLUSTERING PROBLEM

There are two general ways of comparing clustering methods; the first is to consider how easy they are to use, and the second is to evaluate how well they perform when used. The first type is usually computer-oriented and considers execution time and storage requirements (see [1]). In actual application of any clustering method, such a criterion must be considered. Yet how well a clustering method performs is still the ultimate concern. While a quick good method is always better than a slow bad one, the choice between a quick bad method and a slow good one requires some quantification of what makes a method good or bad.

Evaluation of the performance of a clustering method requires some means of comparing its results to either standard results, or to the results of another method. This section develops such a measure of the similarity between clusterings, while the next section proposes several standard situations in which this measure can be used to evaluate methods on the basis of different aspects of their performance.

In standard classification problems, such as discriminant function analysis, there is a correct classification against which to view the results of classification schemes. Often a measure of the "goodness" of such a system is a simple count of the points misclassified or this number normalized into percent error. In the clustering situation organization is sought which arises from the data themselves, and there is no absolute scheme with which to

* William M. Rand is assistant professor of biostatistics, Massachusetts Institute of Technology, Cambridge, Mass. 02139. This article is based in part on the author's doctoral dissertation, submitted to the University of California at Los Angeles and sponsored by PHS grant GM-00049 and NIH grant FR-3. The author wishes to thank Professor A. A. Afifi, UCLA, for valuable suggestions.

measure clusterings. However, there is a natural extension of this idea involving the comparison of two arbitrary clusterings.

Motivation of this measure follows three basic considerations. First, clustering is discrete in the sense that every point is unequivocably assigned to a specific cluster. Second, clusters are defined just as much by those points which they do not contain as by those points which they do contain. Third, all points are of equal importance in the determination of clusterings. While specific examples in which these considerations do not hold can be devised, these three assumptions form the basis for a general clustering problem. From them it follows that a basic unit of comparison between two clusterings is how pairs of points are clustered. If the elements of an individual point-pair are placed together in a cluster in each of the two clusterings, or if they are assigned to different clusters in both clusterings, this represents a similarity between the clusterings, as opposed to the case in which the elements of the point-pair are in the same cluster in one clustering and in different clusters in the other. From this, a measure of the similarity between two clusterings of the same data, Y and Y', can be defined as $c(Y, Y')$ equal to the number of similar assignments of point-pairs normalized by the total number of point-pairs.

Consider the following illustration which calculates c between two clusterings of six points. Let $Y = \{(a, b, c), (d, e, f)\}$ and $Y' = \{(a, b), (c, d, e), (f)\}$, then the point-pairs are tabulated as follows:

Point-pair	*ab*	*ac*	*ad*	*ae*	*af*	*bc*	*bd*	*be*	*bf*	*cd*	*ce*	*cf*	*de*	*df*	*ef*	*Total*
Together in both	*												*			2
Separate in both			*	*	*		*	*				*				7
Mixed		*				*				*	*			*	*	6

The total of nine similarities out of a possible 15 gives $c(Y, Y') = .6$.

More precisely, given N points, $X_1, X_2, \cdots, X_N$, and two clusterings of them $Y = \{Y_1, \cdots, Y_{K_1}\}$ and $Y' = \{Y_1', \cdots, Y_{K_2}\}$, we define

$$c(Y, Y') = \sum_{i<j}^{N} \gamma_{ij} \Big/ \binom{N}{2}, \tag{2.1}$$

where

$$\gamma_{ij} = \begin{cases} 1 \text{ if there exist } k \text{ and } k' \text{ such that both } X_i \text{ and } X_j \text{ are in both } Y_k \text{ and } Y'_{k'} \\ 1 \text{ if there exist } k \text{ and } k' \text{ such that } X_i \text{ is in both } Y_k \text{ and } Y'_{k'} \text{ while } X_j \text{ is in neither } Y_k \text{ or } Y'_{k'} \\ 0 \text{ otherwise} \end{cases}$$

Note that there is a simple computational form for c. Given a pair of clusterings Y and Y' of the same N points, arbitrarily number the clusters in each clustering and let n_{ij} be the number of points simultaneously in the ith cluster of Y and the jth cluster of Y'. Then the similarity between Y and Y' is:

$$c(Y, Y') = \left[\binom{N}{2} - [1/2\{\textstyle\sum_i (\sum_j n_{ij})^2 + \sum_j (\sum_i n_{ij})^2\} - \sum\sum n^2_{ij}]\right] \Big/ \binom{N}{2}. \tag{2.2}$$

There are three fundamental properties of c. First, it is a measure of similarity, ranging from $c=0$ when the two clusterings have no similarities (i.e., when one consists of a single cluster and the other only of clusters containing single points), to $c=1$ when the clusterings are in fact identical. Second, while c is a measure of similarity, $1-c$ is a measure of distance, being a metric on the set of all clusterings of a given set of points, $\mathcal{Y}$. Third, if a distribution is assumed for X, and under various conditions to be discussed in the next section, c is a random variable.

Since operational clustering methods search a subset of $\mathcal{Y}$ which is usually defined as all of a certain type of rearrangement of a specific initial clustering (such as all clusterings formed by joining pairs of clusters of the original clustering), it is of interest to examine the behavior of the measure c in some of these situations. Table 1 displays several comparisons between similar clusterings and their limits as the number of points and clusters increase. Given an initial clustering Y having k clusters with n points in each, Table 1A considers the similarity between this clustering and new clusterings formed from Y by various simple operations. Table 1B considers the similarity between two clusterings each of which is formed by the same operation from the same initial clustering but with the operation applied to different parts of the original clustering. Table 1C shows the similarity of the initial clustering to certain standard clusterings, and to the clustering which can be considered the opposite clustering, that which consists of n clusters each of which contains k points, one from each of the original k clusters. Note that in all three parts of the table the similarities depend on both k, the number of clusters, and n, the number of points in each cluster. This follows directly from the fact that the similarity of two clusterings is essentially the proportion of point-pairs whose relationship is the same in both. Thus the joining of two of three clusters is much more significant than the joining of two of thirty (given equal sized clusters).

3. EVALUATION OF METHODS

The most important problem facing an investigator with data he would like to examine by clustering methods

1. EXPRESSIONS FOR THE MEASURE c BETWEEN TWO SIMILAR CLUSTERINGS, GIVEN AN INITIAL CLUSTERING, Y, WHICH HAS k CLUSTERS OF n POINTS EACH

A. c(Y,Y'), where Y' is a simple modification of Y

Modification of Y	c(Y,Y')	Limit of c: $n \to \infty$	Limit of c: $k \to \infty$
Two clusters joined	$\frac{(k^2-2)n-k}{k^2n-k}$	$\frac{k^2-2}{k^2}$	1
One cluster split into two equal parts (n even)	$\frac{(2k^2-1)n-2k}{2k^2n-2k}$	$\frac{2k^2-1}{2k^2}$	1
One cluster split into single point clusters	$\frac{(k^2-1)n-k+1}{k^2n-k}$	$\frac{k^2-1}{k^2}$	1
One point taken from each cluster to form a new cluster of k points	$\frac{kn^2-3n-k+3}{kn^2-n}$	1	$\frac{n^2-1}{n^2}$

B. c(Y',Y"), where Y' and Y" are similar modifications of the original clustering Y

Differences between Y' and Y"	c(Y',Y")	Limit of c: $n \to \infty$	Limit of c: $k \to \infty$
Movement of a point to different clusters	$\frac{k^2n-k-4}{k^2n-k}$	1	1
Different clusters split into two equal parts (n even)	$\frac{(k^2-1)n-k}{k^2n-k}$	$\frac{k^2-1}{k^2}$	1
Different pairs of clusters joined	$\frac{(k^2-4)n-k}{k^2n-k}$	$\frac{k^2-4}{k^2}$	1

C. c(Y,Y'), where Y' is a major modification of the original clustering Y

Modification of Y	c(Y,Y')	Limit of c: $n \to \infty$	Limit of c: $k \to \infty$
All clusters joined into one large cluster	$\frac{n-1}{kn-1}$	$\frac{1}{k}$	0
All clusters split into single point clusters	$\frac{(k-1)n}{kn-1}$	$\frac{k-1}{k}$	1
n clusters are formed with k points in each, one point from each original cluster	$\frac{(k-1)(n-1)}{kn-1}$	$\frac{k-1}{k}$	$\frac{n-1}{n}$

is that of which method to use. While many researchers would be unable to specify the specific method which best suited their needs, most could suggest characteristics such a method should possess. The preceding machinery allows clustering methods to be evaluated with respect to such requirements. The following four questions illustrate this process of evaluation for four fundamental aspects of clustering methods. Each is translated into a distribution problem which is investigated in the next section by Monte Carlo techniques for two specific clustering methods.

3.1 How well does a method retrieve "natural" clusters?

Clustering methods produce clusterings irrespective of the presence or absence of 'natural' structure within the data. An important consideration is what happens when there is obvious structure present. Consider N points drawn randomly from K distinct distributions (differing only in location) with $n_i \geq 1$ points from the ith distribution and denote as Y the clustering which clusters together points which are drawn from the same distribution. A specific method applied to these data produces a clustering, Y'. For each sample of N points $c\,(Y, Y')$ can be calculated; its distribution represents how well the method retrieves clusters inherent in the data.

3.2 How sensitive is a method to perturbation of the data?

In many applications it is not known whether the data are good representations of their respective populations. The changes of clustering which result from slight movement of points are therefore of critical importance in both choice of methods and interpretation of results. Consider N points drawn randomly from a specific distribution and clustered by a specific method as Y. Add to each point a quantity drawn from a distribution with zero mean and small variance and cluster these perturbed points by the same method to produce Y'. The distribution of $c(Y, Y')$ indicates the sensitivity of the particular method to errors of measurement or resampling.

3.3 How sensitive is a method to missing individuals?

Sometimes an investigator knows, or suspects, that his data set is incomplete, that whole subpopulations are missing or not well represented. In this case he is interested in the agreement or lack of agreement between the clusterings he derives from the data he has and the clusterings he would get if he had more data. Consider N_1 points drawn from a single distribution and clustered by a specific method as Y. If N_2 additional points are drawn from the same distribution and all N_1+N_2 points are clustered by the same method as Y', the assignments of the original N_1 points in Y' define the clustering Y''. Assuming that Y'' represents how the initial N_1 points should have been clustered, $c(Y, Y'')$ describes how close the specific method comes to finding this clustering using only those N_1 points.

3.4 Given two methods, do they produce different results when applied to the same data?

An investigator, trying to choose between methods, would be helped by knowing how different the methods are in terms of the clusterings they produce. Given, for example, a complex method which requires much computing, it would be of value to determine if a simpler method could be used as an approximation. (This question also suggests a stopping criteria for iterative methods. Thus, a method could be iterated until results of successive steps agreed within a prechosen similarity.) Consider N points drawn from a given distribution. Clustering by one method as Y and by another as Y' permits the agreement between the two methods for any specific number of clusters to be measured as $c(Y, Y')$.

4. EXAMPLES OF USAGE

These four questions were formulated as procedures for the evaluation of methods (details are given later in this section) and applied to two simple clustering meth-

2. RETRIEVAL: COMPARISON OF THE ABILITY OF TWO CLUSTERING METHODS TO RETRIEVE FIVE MULTIVARIATE NORMAL POPULATIONS

Number of clusters	Average of c		Standard deviation of c		Percentage of complete agreement	
	Method					
	T/N	AA	T/N	AA	T/N	AA
10	.87	.71	.020	.018	0	0
9	.88	.68	.023	.019	0	0
8	.88	.68	.027	.016	0	0
7	.89	.63	.034	.015	0	0
6	.89	.56	.039	.013	0	0
5	.88	.50	.046	.013	3	0
4	.84	.43	.038	.013	0	0
3	.75	.35	.045	.007	0	0
2	.60	.26	.048	.004	0	0

ods. These methods are both agglomerative in that, given a best clustering of K clusters, they examine all clusterings of $K-1$ clusters formed by joining pairs of clusters, and select one for which their specific criterion is a minimum. These methods were applied in a stepwise fashion, starting with the clustering in which each point is itself a cluster and proceeding until all points are in a single cluster. Thus for each set of points a sequence of best clusterings is generated. The important question of how to choose which is the best number of clusters is not considered here.

The first clustering method, denoted by T/N, takes for its criterion the sum of all within distances (those distances between points which are in the same cluster) divided by the number of such distances. Denoting the total of the distances between the n_k points in the kth cluster as W_k, this criterion is written $\sum W_k / \sum \binom{n_k}{2}$, where the summation is over all K clusters. The second method, denoted by AA, takes as its criterion the average of the average within distance, or $1/K \sum W_k / \binom{n_k}{2}$. These methods were chosen for their simplicity and similarity, since both minimize a type of within distance.

The procedures were applied by means of Monte Carlo simulation, each being replicated 100 times. The distribution of c is described by three statistics, the mean, the standard deviation, and the percentage of complete agreement. Sample results for $K=2, 3, \cdots, 10$ are displayed in Tables 2 through 5.

For retrieval (Section 3.1), six points were drawn randomly from each of five five-dimensional normal populations with unity covariance matrix and means symmetrically four units apart. These points were clustered by each of the two methods to produce sequences of clusterings. These clusterings were then compared (c calculated) with the clustering which clustered together those points drawn from the same population. Table 2 displays the results of the application of this procedure.

For perturbation (Section 3.2), 30 points were drawn randomly from a five-dimensional unit normal distribution and a clustering sequence produced by application of each method. Random perturbations drawn from $N(0, .01)$ were added to each coordinate of each of the 30 points and a new clustered sequence derived. These sequences were compared for each K to produce Table 3.

For missing data (Section 3.3), 25 points were drawn from a five-dimensional unit normal distribution and clustered. Then an additional five points were drawn from the same distribution and the total 30 points clustered. The assignments of the original 25 points within this sequence of clusterings were then compared with the clustering sequence based on only the original 25 for each K. The results are summarized in Table 4.

For comparison of the two methods (Section 3.4), 30 points were randomly chosen from a five-dimensional unit normal distribution. Each method was applied to produce a clustering sequence and the sequences compared for each value of K (see Table 5).

The application of these procedures illustrates their utility. Table 5 shows that the examined methods are not similar, while the other three tables indicate where the dissimilarities lie. Essentially, method T/N is better for retrieval of structure while method AA is less affected by missing data. The situation with regard to perturbation illustrates the further generalization that method AA

3. PERTURBATION: COMPARISON OF THE SENSITIVITY OF TWO CLUSTERING METHODS TO SLIGHT MOVEMENT OF OBJECTS BEING CLUSTERED[a]

Number of clusters	Average of c		Standard deviation of c		Percentage of complete agreement	
	Method					
	T/N	AA	T/N	AA	T/N	AA
10	.92	.68	.080	.022	0	0
9	.91	.65	.082	.022	0	0
8	.89	.65	.082	.028	0	0
7	.87	.61	.091	.035	0	0
6	.85	.59	.110	.041	0	0
5	.81	.60	.116	.046	0	0
4	.76	.63	.121	.060	0	1
3	.68	.70	.106	.075	0	2
2	.61	.81	.091	.132	0	17

[a] The similarity of the clusterings of the data before and after perturbation is measured by c.

4. MISSING DATA: COMPARISON OF TWO CLUSTERING METHODS IN TERMS OF THE EFFECT ELIMINATION OF OBJECTS HAS ON THE ASSIGNMENT OF THOSE LEFT[a]

Number of clusters	Average of c		Standard deviation of c		Percentage of complete agreement	
	Method					
	T/N	AA	T/N	AA	T/N	AA
10	.96	.84	.020	.102	6	6
9	.95	.82	.023	.103	2	4
8	.94	.80	.030	.106	3	0
7	.93	.77	.032	.117	1	1
6	.91	.75	.039	.119	1	2
5	.87	.78	.054	.133	1	16
4	.82	.81	.071	.131	2	24
3	.76	.88	.103	.114	3	42
2	.68	.94	.163	.104	7	72

[a] The similarity of clusterings with and without these additional objects is measured by c.

5. DIRECT COMPARISON OF THE AGREEMENT BETWEEN TWO CLUSTERING METHODS, T/N AND AA, WHEN APPLIED TO SAME SETS OF DATA

Number of clusters	Average of c	Standard deviation of c	Percentage of complete agreement
10	.76	.015	0
9	.72	.018	0
8	.71	.022	0
7	.64	.019	0
6	.57	.023	0
5	.50	.025	0
4	.44	.027	0
3	.43	.032	0
2	.52	.041	0

is better for small K while method T/N is better for larger K.

REFERENCES

[1] Ball, Geoffrey H., "Data Analysis in the Social Sciences: What About the Details?" in *AFIPS Conference Proceedings*, Vol. 27, Part 1: *Fall Joint Computer Conference*, (1965), 533–59.

[2] Fisher, Walter D., "On Grouping for Maximum Homogeneity," *Journal of the American Statistical Association*, 53 (December 1958), 789–98.

[3] Fortier, J. J. and Solomon, H., "Clustering Procedures," in P. R. Krishnaiah, ed., *Multivariate Analysis*, New York: Academic Press, 1966, 493–506.

[4] Friedman, H. P., and Rubin, J., "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association*, 62 (December 1967), 1159–78.

[5] Jardine, J. and Sibson, R., "The Construction of Hierarchic and Non-hierarchic Classifications," *The Computer Journal*, 11 (August 1968), 177–84.

[6] Lance, G. N. and Williams, W. T., "A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems," *The Computer Journal*, 9 (February 1967), 373–80.

[7] ———, and Williams, W. T., "A General Theory of Classificatory Sorting Strategies. II. Clustering Systems," *The Computer Journal*, 10 (November 1967), 271–7.

[8] Rubin, J., "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem," *Journal of Theoretical Biology*, 15 (April 1967), 103–44.

[9] Ward, Joe H., Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58 (March 1963), 236–44.