[2] K. S. Fu, *Syntactic Methods in Pattern Recognition.* New York: Academic, 1974.
[3] R. L. Kashyap, "Pattern recognition and database," in *Proc. Pattern Recognition and Image Processing Conf.*, Troy, NY, June 1977.
[4] S. A. Boorman and D. C. Oliver, "Metrics on spaces of finite trees," *J. Math. Psych.*, vol. 10, Oct. 1973.
[5] R. A. Wagner and M. J. Fisher, "The string-to-string correction problem," *J. Ass. Comput. Mach.*, vol. 21, Jan. 1974.
[6] W. S. Brainerd, "Tree generating regular systems," *Inform. Contr.*, vol. 14, 1969.
[7] A. V. Aho and T. G. Peterson, "A minimum distance error-correcting parser for context-free languages," *SIAM J. Comput.*, vol. 4, Dec. 1972.
[8] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.
[9] K. S. Fu and S. Y. Lu, "A sentence-to-sentence clustering procedure for pattern analysis," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, May 1978.
[10] S. Y. Lu and K. S. Fu, "Error-correcting tree automata for pattern recognition," *IEEE Trans. Comput.*, vol. C-27, Nov. 1978.

A Cluster Separation Measure

DAVID L. DAVIES AND DONALD W. BOULDIN

***Abstract*—A measure is presented which indicates the similarity of clusters which are assumed to have a data density which is a decreasing function of distance from a vector characteristic of the cluster.**

The measure can be used to infer the appropriateness of data partitions and can therefore be used to compare relative appropriateness of various divisions of the data. The measure does not depend on either the number of clusters analyzed nor the method of partitioning of the data and can be used to guide a cluster seeking algorithm.

***Index Terms*—Cluster, data partitions, multidimensional data analysis, parametric clustering, partitions, similarity measure.**

INTRODUCTION

Although many clustering systems depending on distance criteria have been developed [1]-[5], a recurrent and largely unsolved problem has been the determination of the proper number of clusters in data. There are two principal approaches to this problem. One commonly used technique depends on plotting an optimization parameter against a number of clusters and choosing as optimum a cluster number for which a large change occurs in the parameter value. Various parameters have been suggested as the performance index [6], [7]. A second method is hierarchical. Hierarchical techniques generally seek large changes in intergroup fusions. Fusion distances are generally determined with the aid of a dendrogram [1], [2]. Both of these techniques depend on the determination of relatively large changes in an index rather than its minimization or maximization, and therefore, in general, require human interpretation and subjective analysis of what is to be considered a "large change" in the parameter.

In this correspondence a new clustering parameter is presented. The minimization of this parameter appears to indicate natural partitions of data sets.

The cluster separation measure incorporates the fundamental features of some of the well-accepted similarity measures often applied to the cluster analysis problem and also satisfies certain heuristic criteria.

Manuscript received April 3, 1978; revised September 14, 1978. This work was supported in part by the Defense Advanced Research Projects Agency/Space and Missile Systems Organization under Contract F04701-77-C-0072.

D. L. Davies was with the Department of Electrical Engineering, University of Tennessee, Knoxville, TN 37916. He is now at 17 C Downey Drive, Manchester, CT 06040.

D. W. Bouldin is with the Department of Electrical Engineering, University of Tennessee, Knoxville, TN 37916.

THEORETICAL FORMULATION

It was decided that a general cluster separation measure should possess the following attributes.

1) It should require little or no user interaction or specification of parameters.
2) It should be applicable to hierarchical data sets.
3) It should be computationally feasible for relatively large data sets.
4) It should yield meaningful results for data of arbitrary dimensionality.

With these attributes in mind, the following definitions are made.

Definition 1: A real-valued function is said to be a distance function or metric if the following properties hold:

1) $d(X_i, X_j) \geq 0 \qquad \forall X_i, X_j \in E_p$
2) $d(X_i, X_j) = 0 \qquad \text{iff} \quad X_i = X_j$
3) $d(X_i, X_j) = d(X_j, X_i) \qquad \forall X_i, X_j \in E_p$
4) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j) \qquad \forall X_i, X_j, X_k \in E_p$

where E_p is a p-dimensional Euclidean space [9].

Definition 2: A real-valued function is said to be a dispersion measure if the following properties hold: let cluster C have members $X_1, X_2, \cdots, X_m \in E_p$

1) $S(X_1, X_2, \cdots, X_m) \geq 0$
2) $S(X_1, X_2, \cdots, X_m) = 0 \quad \text{iff} \quad X_i = X_j \; \forall X_i, X_j \in C.$

The goal is to define a general cluster separation measure, $R(S_i, S_j, M_{ij})$, which allows computation of the average similarity of each cluster with its most similar cluster. With that in mind, let us define a cluster similarity measure as follows.

Definition 3: A real-valued function is a cluster similarity measure if the following properties hold:

1) $R(S_i, S_j, M_{ij}) \geq 0,$
2) $R(S_i, S_j, M_{ij}) = R(S_j, S_i, M_{ji}),$
3) $R(S_i, S_j, M_{ij}) = 0 \quad \text{iff} \quad S_i = S_j = 0,$
4) if $S_j = S_k$ and $M_{ij} < M_{ik}$
   then $R(S_i, S_j, M_{ij}) > R(S_i, S_k, M_{ik}),$
5) if $M_{ij} = M_{ik}$ and $S_j > S_k$
   then $R(S_i, S_j, M_{ij}) > R(S_i, S_k, M_{ik}),$

where M_{ij} is the distance between vectors which are chosen as characteristic of clusters i and j, and S_i and S_j are the dispersions of clusters i and j, respectively.

Definition 3 imposes certain limitations on R which are somewhat arbitrary but, nevertheless, heuristically meaningful. It indicates that

1) the similarity function R is nonnegative,
2) it has the property of symmetry,
3) the similarity between clusters is zero only if their dispersion functions vanish,
4) if the distance between clusters increases while their dispersions remain constant, the similarity of the clusters decreases,

5) if the distance between clusters remains constant while the dispersions increase, the similarity increases.

The conditions in Definitions 2 and 3 are minimal necessary conditions. The above properties suggest that R be formed using some functions $F(S_i, S_j)$ and $G(M_{ij})$ in a reciprocal relationship.

The function below is offered as one which satisfies the required criteria and which reduces to certain familiar similarity measures for special choices of dispersion measures, distance measures, and characteristic vectors.

Definition 4:

$$R_{ij} \equiv \frac{S_i + S_j}{M_{ij}}$$

with S_i, S_j and M_{ij} as defined above. It is clear that R_{ij} is one of the simplest functions which satisfies Definition 3. $\bar{R}$ is then defined as:

Definition 5:

$$\bar{R} \equiv \frac{1}{N}\sum_{i=1}^{N} R_i$$

where $R_i \equiv$ maximum of R_{ij} $i \neq j$.

$\bar{R}$ has the significance of being the system-wide average of the similarity measures of each cluster with its most similar cluster. The "best" choice of clusters, then, will be that which minimizes this average similarity. In order to demonstrate the use of $\bar{R}$, the following distance function, dispersion measure, and characteristic vector were chosen:

$$S_i = \left\{\frac{1}{T_i}\sum_{j=1}^{T_i} |X_j - A_i|^q\right\}^{1/q}$$

where T_i is the number of vectors in cluster i.

A_i is the centroid of cluster i

$$M_{ij} = \left\{\sum_{k=1}^{N} |a_{ki} - a_{kj}|^p\right\}^{1/p}$$

where a_{ki} is the kth component of the n-dimensional vector a_i, which is the centroid of cluster i.

It should be noted that M_{ij} is the Minkowski metric [6] of the centroids which characterize clusters i and j. When $p = 1$, M_{ij} reduces to the "city block" distance used by Carmichael and Sneath [2]. When $p = 2$, M_{ij} is the Euclidean distance between centroids. S_i is the qth root of the qth moment of the points in cluster i about their mean. If $q = 1$, S_i becomes the average Euclidean distance of vectors in cluster i to the centroid of cluster i. If $q = 2$, S_i is the standard deviation of the distance of samples in a cluster to the respective cluster center. If $p = q = 2$, R_{ij} is the reciprocal of the classic Fisher similarity measure calculated for clusters i and j.

Use of the cluster measure $\bar{R}$ does not depend on a particular clustering algorithm. It can be used to compare the validity of data partitions regardless of how those partitions were generated. It can be used to compare partitions with either similar or different numbers of clusters.

As a simple illustration of the use of $\bar{R}$ to compare different partitions which have the same number of clusters, consider the four point data set shown in Fig. 1. It is desired to partition the data into two sets using a K-means algorithm similar to that described in Tou and Gonzalez [1]. If points (1, 1) and (1, 3) are chosen as initial centers, the algorithm will produce the partition indicated by the indicated Surface 1, with cluster centers at (3, 1) and (3, 3). If points (1, 1) and (5, 1) are chosen as initial centers, the algorithm will separate the data as indicated by partition Surface 2 with cluster centers at (1, 2) and (5, 2). If we calculate $\bar{R}$ for each of the partitions

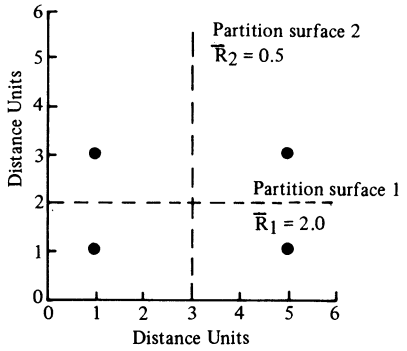


Fig. 1. Comparison of two partitions of 4 point data set and use of $\bar{R}$ to evaluate their relative appropriateness.

with $p = q = 2$, $\bar{R}_1 = 2.0$ for the first partition, and for the second partition $\bar{R}_2 = 0.5$. Since $\bar{R}_2 < \bar{R}_1$, the second partition is taken to be the superior of the two.

In fact, one can infer from the value of $\bar{R}_1$, that it reflects a particularly inappropriate partition. Random two-dimensional data yield minimum $\bar{R}$ values of approximately 0.6 if single member clusters are prohibited. A value of $\bar{R}$ above, or in the same range as the minima obtained for randomly distributed data, indicates that a particular partition does not separate data into natural clusters.

It is appropriate to mention that a data set must be partitioned into at least two groups with different cluster centers for $\bar{R}$ to have meaning. This is a mathematical necessity since the distance measure in the denominator of $\bar{R}$ must be nonzero for it to be defined. The use of $\bar{R}$ also becomes limited if clusters containing single members are permitted, since such clusters have zero dispersion (according to Definition 2, property 2). If $\bar{R}$ is to be used as a parameter to guide a cluster seeking algorithm, these two limitations should be kept in mind. For the following demonstration it was decided to require that each cluster contain at least two (not necessarily unique) members. Neither the chosen cluster seeking algorithm nor the decision as to how single member clusters are handled should be taken as unique or as optimum. $\bar{R}$ could be used to evaluate the appropriateness of data partitions generated using other algorithms such as those described in Duda and Hart [12] and Tou and Gonzalez. [1] Similarly, one might decide to deal with the single member cluster problem differently, for example, by permitting a certain number of clusters to have only one member. Clearly, if an unlimited number of single member clusters is permitted, $\bar{R}$ will have a minimum value of zero when each unique data point forms its own cluster.

Implementation and Results

Demonstration of the measure $\bar{R}$ will now be given on a variety of data sets. For each of the analyses presented, p was chosen equal to 2, which means that M_{ij} became the Euclidean distance between the centroids of clusters i and j.

A K-means algorithm was employed to determine K cluster centers given K seed centers [1]. A value of $\bar{R}(K)$ was then calculated. The cluster center pair which has the largest contribution to $\bar{R}$ (i.e., the most similar cluster pair) had its two centers replaced by a single center located at the centroid of the set formed by the union of the cluster pair. Thus, K was reduced by one, and the process was iterated. This process terminated when only two centers remained, since a cluster similarity measure has no meaning for a single cluster.

The initial number of clusters was chosen equal to the number of vectors in a given data set. Since it was decided (heuristically) that a single point could not comprise an interesting cluster, an arbitrarily large dispersion was assigned clusters with single

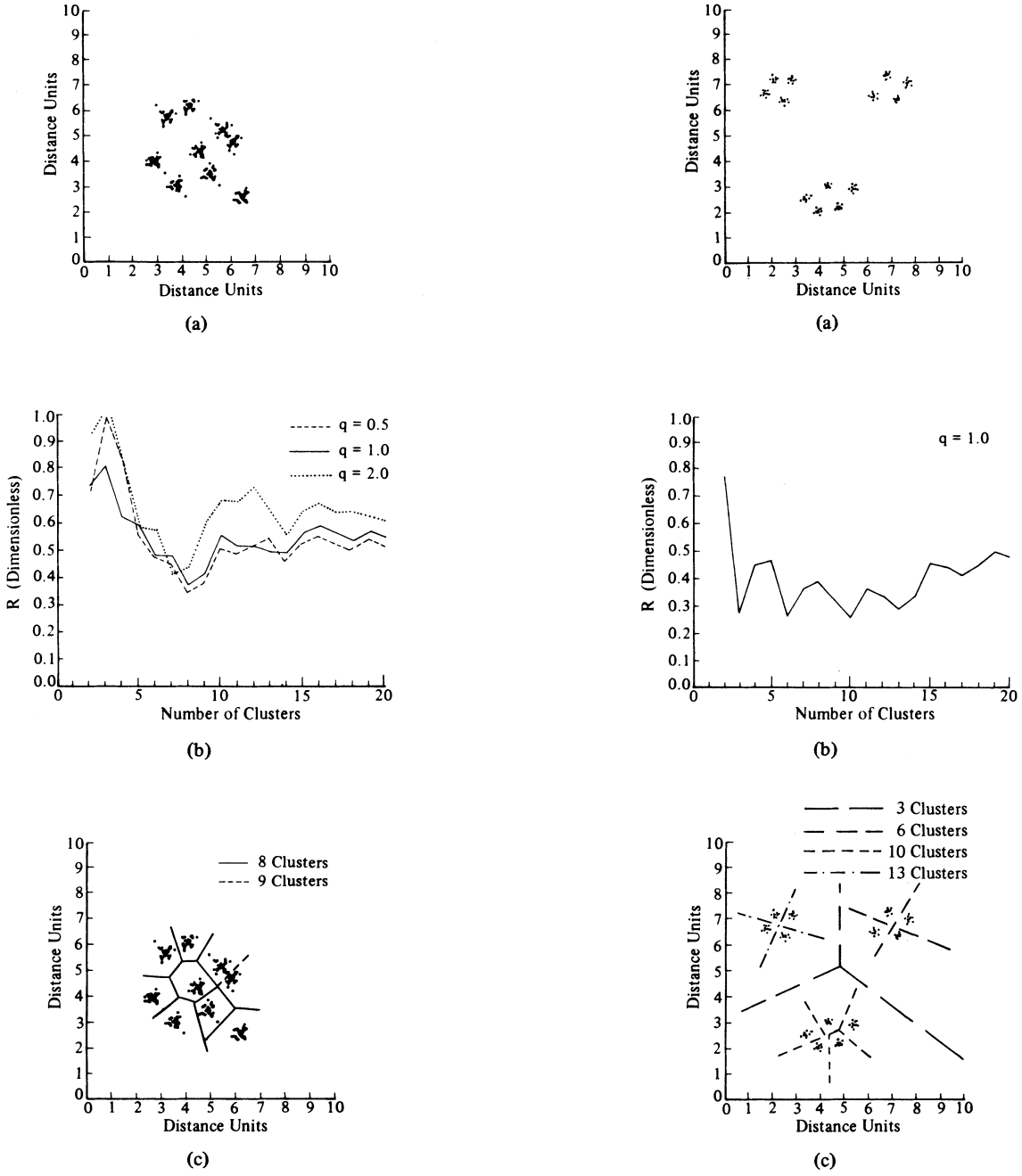


Fig. 2. (a) ISODATA data set. (b) Cluster separation measure graph for ISODATA data. (c) Separation of ISODATA data set into eight and nine clusters.

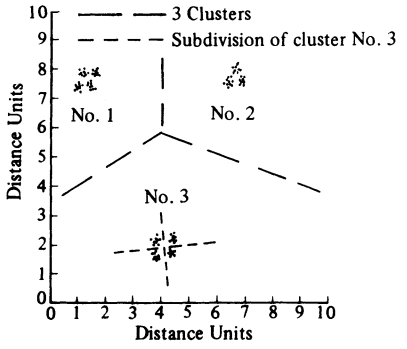
Fig. 3. (a) Hierarchical data set. (b) Cluster separation measure graph for Fig. 3(a). (c) Separation into three, six, ten, and thirteen clusters.

membership. Thus, all single member clusters were eliminated before any which had multiple membership. The separation measure was tested on data sets of dimensionality 1, 2, and 4 with the major tests performed on two-dimensional data for ease of visual presentation.

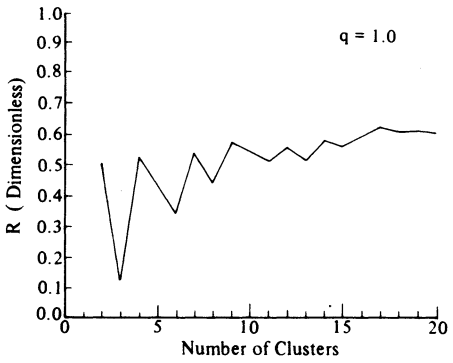
Fig. 2(a) shows a data set of 225 points adapted from Ball and Hall's [7] (ISODATA) cluster system test set. The associated graph of Fig. 2(b) shows the performance of $\overline{R}$ for the smallest 20 values of K for $q = 0.5$, 1.0, 2.0 and $p = 2.0$. Ball and Hall generated the data using nine approximately Gaussian, parent distributions. Since we transferred the data using a digitization tablet, minor discrepancies in the data were introduced. As shown in Fig. 2(b), $\overline{R}$ is a minimum for $K = 8$, with $K = 9$ producing a value of $\overline{R}$ approximately 10 percent greater. The associated distribution of data points into eight clusters is indicated by continuous lines with the ninth indicated by a dashed line in Fig. 2(c).

As a second two-dimensional example, the use of $\overline{R}$ is demonstrated in the analysis of a hierarchical system. A data set of 130 points and its associated $\overline{R}$ graph is shown in Fig. 3(a) and 3(b). The partitions corresponding to the four local minima of $\overline{R}$ are indicated in Fig. 3(c). That the local minima for $\overline{R}$ at $K = 3$, 6, 10, and 13 are approximately equal is a result of the roughly equivalent densities in each of the 13 small clusters and their approximately equal separation in the large clusters. The cluster separation measure indicates that adjacent small clusters have approximately the same similarity as do the large clusters.

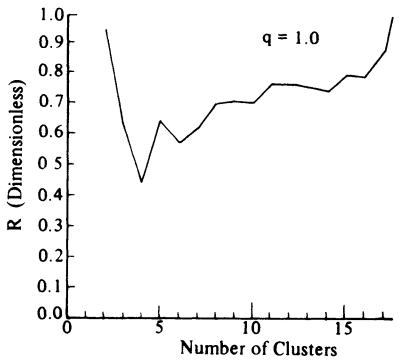
Another example of hierarchical data is given in Fig. 4. The 110 data points and the associated $\overline{R}$ graph are shown in



(a)



(b)



(c)

Fig. 4. (a) Second hierarchical data set. (b) Cluster separation measure graph for Fig. 4(a). (c) Cluster separation measure for data in cluster number three of Fig. 4(a).

Fig. 4(a) and 4(b) with the three group partition chosen as optimum indicated by dashed lines. A subsequent analysis is carried out on those points assigned to group number three. The associated $\overline{R}$ graph is shown in Fig. 4(c), while the separation indicated as optimum is shown by the fine dotted lines in Fig. 4(a). It is interesting to note that the minimum $\overline{R}$ in Fig. 4(c) is above the minimum $\overline{R}$ in Fig. 4(b). If this were not the case, cluster three in Fig. 4(a) would have been subdivided in the optimum division of the global data set.

As a final example, $\overline{R}$ was calculated for the four-dimensional iris data used by Fisher in a series of statistical tests [8]. Four measurements were made on each of fifty members of each of the three varieties of flower Iris Setosa, Iris Versicolor, and Iris Virginica. Iris Setosa is known to be linearly separable from

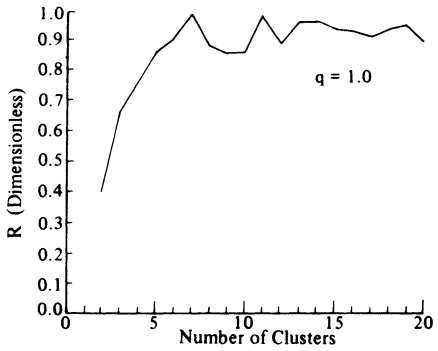


Fig. 5. Cluster separation measure graph of IRIS data.

the other two, while there is considerable overlap between Iris Virginica and Iris Versicolor. As shown in Fig. 5, $K = 2$ was chosen as the best division of data while local minima for $\overline{R}$ are found at $K = 9$ and $K = 17$. Both the numbers of clusters and their associated data separations correlate well with the results obtained by other workers [11] using more involved methods.

Conclusion

Cluster analysis is often one of the first steps in the analysis of data. As such, it is an effort at unsupervised learning usually in the context of very little *a priori* knowledge. Therefore, the requirement that a user supply an analysis system with parameter values, such as minimum acceptable cluster distance or minimum acceptable standard deviation, knowledge of which presumes previous study of the data, is a major detriment of such systems. In fact, as Chen [10, p. 135] points out, "a common drawback to all cluster algorithms is that their performance is highly dependent on the user setting various parameters. In fact, the "proper" setting usually can only be determined by a trial and error method." If incorporated into a cluster seeking algorithm, the measure presented here substantially overcomes this difficulty by requiring the user to specify only the p and q exponents, which is equivalent to requiring the user to specify only the distance and dispersion measures to be used.

References

[1] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.
[2] B. Everett, *Cluster Analysis*. New York: Wiley, 1975.
[3] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.
[4] R. M. Haralick, "Automatic remote sensor image processing," in *Topics in Applied Physics, Digital Picture Analysis*. New York: Springer-Verlag, 1976.
[5] C. H. Chen, "Theory and applications of imagery pattern recognition," in *Proc. 4th Int. Congr. Sterology*, National Bureau of Standards Publ. 431, Washington, DC, 1976.
[6] H. P. Friedman and J. Rubin. "On some invariant criteria for Grouping Data," *J. Amer. Stat. Assoc.*, vol. 62, pp. 1159-1178, 1967.
[7] R. L. Thorndike, "What belongs in a family?" *Psychometrika*, vol. 18, pp. 267-276, 1953.
[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," in *Machine Recognition of Patterns*, A. K. Agrawala, Ed. New York: IEEE Press, 1976.
[9] B. S. Duran and P. L. Odell, *Cluster Analysis*. Berlin: Springer-Verlag, 1974.
[10] C. H. Chen, *Statistical Pattern Recognition*. Rochelle Park, NJ: Hayden, 1973.
[11] J. K. Bryan, *Classification and Clustering Using Density Estimation*, Ph.D. dissertation, Univ. of Missouri, Columbia, MO, 1971.
[12] R. D. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.