DATA ANALYSIS AND INFORMATICS

Proceedings of the Second International Symposium on
Data Analysis and Informatics,
organised by the Institut de Recherche d'Informatique et d'Automatique,
Versailles, October 17-19, 1979

edited by

E. DIDAY
University Paris IX - IRIA

L. LEBART
Centre National de la Recherche Scientifique - CREDOC

J. P. PAGÈS
Commissariat à l'Energie Atomique

R. TOMASSONE
Institut National de la Recherche Agronomique - CNRZ

1980

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD

Data Analysis and Informatics
E. Diday et al. (eds.)
© North-Holland Publishing Company, 1980

EVALUATION PROCEDURE OF AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS BY FUZZY RELATIONS

Noboru Ohsumi
The Institute of Statistical Mathematics
Tokyo, Japan

This paper will be mainly concerned with the evaluation of techniques of cluster analysis, in particular agglomerative hierarchical clustering (AHC) methods. In the AHC methods, relationships among the objects being grouped are represented by a dissimilarity or similarity matrix. Therefore it is quite natural and meaningful to describe the matrix by taking it as a representation of a relation or a graph. In such cases, especially, the concept of fuzzy relations proposed by Zadeh is more relevant and useful for a reasonable examination of the clustering models.

We shall first attempt to examine several properties of the AHC methods based on the fuzzy theory, especially single linkage and complete linkage. Furthermore, we shall propose a fuzzy degree of fitness which is a new index for evaluating and comparing the relationship between two relations. This index may be generated by using the fuzzy symmetric difference between two relations. Finally, we shall consider the comparison between the sets of partitions formed by clustering process.

1. INTRODUCTION

In this paper, we shall discuss mainly the properties of the agglomerative hierarchical clustering (AHC) methods. Most AHC methods start the clustering process by forming a matrix which represents the pairwise similarities or dissimilarities of all objects being classified. In general, the solution of an AHC method can be represented by a hierarchical structure, that is, a hierarchical tree or a dendrogram. But it is rare that the hierarchical structure or dendrogram is constructed explicitly by fusing of the objects. Therefore, the AHC methods can be interpreted as the result of successive approximations in the formation of a hierarchical structure from the original similarity or dissimilarity matrix which represents the kind of relationship between the objects. There are a large number of well-known AHC methods, for example, single linkage, complete linkage, the centroid method, the group average method, Ward's method, and so on.

Though the investigations for evaluation or comparison of the properties of these methods are only little discussed systematically in works, it is quite important and necessary to discuss them in order to enable us to construct medels with a satisfactory validity from the clustering process.

However, since techniques such as cluster analysis are more interdisciplinary, many concepts are similar. In addition, when we examine the techniques proposed by many researchers in the distinct fields, for instance, biology and psychology, we can observe that most of those are the same or almost the same methods. For example, the researchers refering to Johnson's paper in psychology use the terms "maximum method" and "minimum method", but these two methods are known as "complete linkage" and "single linkage" in the biological field. The terms "complete linkage", "furthest neighbor", "rank order typal analysis", "diameter analysis" are synonyms. The terms "single linkage", "nearest neighbor", "minimum method", "elementary linkage analysis", "connected method" and a kind of minimum spanning tree are synonyms.

Thus turning our attention to the fact that different terms have been used to describe the same thing, we attempt to introduce the fuzzy set theory and the fuzzy relation into the systematic consideration of cluster analysis, especially AHC methods.

2. RELATIONSHIP BETWEEN THE AHC METHODS AND FUZZY RELATIONS

Since the investigations for evaluation or comparison of AHC methods have been only little discussed in formal works up to the present, it is really important and necessary to discuss them in order to study cluster analysis. Therefore, we shall turn our attention to these problems. Solving these problems appears to require some adaptive tools. In such cases, fortunately, it seemed to us that the concept of fuzzy relations proposed by Zadeh is more useful for examining the clustering models. We shall now attempt to examine the several characteristics of the AHC methods and the relationship between them and fuzzy relations. For simplicity in our discussion, we shall define several notations and terms.

We define the set of n objects

$$E = \{ O_1, O_2, O_3, \ldots, O_i, \ldots, O_n \}$$

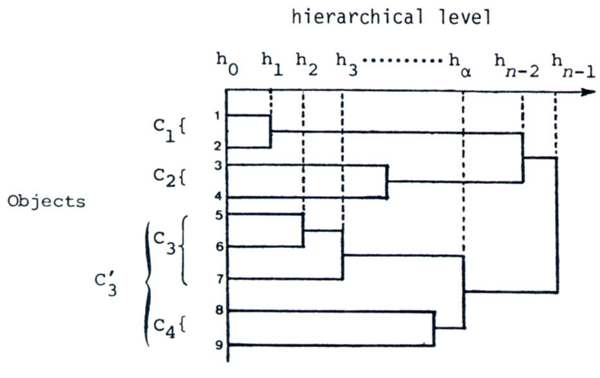
or for abbreviation,

$$E = \{ 1, 2, 3, \ldots, i, \ldots, n \}$$

and denote the raw data consisting of a $n \times m$ matrix.

$$X = (x_{il}) \quad (i=1,2,3,\ldots,n \; ; \; l=1,2,3,\ldots,m)$$

where $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'$ is the observed vector for the ith objects. Then, the AHC methods begin with the computation of a similarity matrix $S=(s_{ij})$ or a dissimilarity matrix $D=(d_{ij})$ between the objects formed from X. Most AHC methods are commonly suitable for using various kinds of dissimilarity or similarity, whether d_{ij} and s_{ij} are metric or non-metric. In the AHC methods, the goal of a clustering process can be represented as a dendrogram. In other words, the input is a matrix D or S, the end of a clustering process is a dendrogram which is a graphical representation of hierarchical structure. The hierarchical structure or dendrogram may be represented by a tree diagram as shown in Figure 1, which is a two dimensional diagram configurating the fusions between objects which have been constructed at each successive level, namely, the hierarchical level or index h_α ($\alpha=0,1,2,\ldots,n-1$). As shown in Figure 1, the order of fusion level is monotonically changing (increasing or decreasing), that is, the hierarchical structure possesses the property of monotone transformation.



This example consists of the nine objects. And there exists the set of cluster $C^6 = \{C_1, C_2, C'_3\}$ at the hierarchical level $h_\alpha (=h_6)$.

Figure 1. A dendrogram with monotonic invariant property.

In the following discussion, we treat only the AHC methods in which the result of clustering may be represented by the monotonic hierarchical structure H.

Thus a dendrogram is considered to be a hierarchical structure H specified by the hierarchical index h(·). And we shall denote such a dendrogram by <H,h>. For example, Figure 1 shows a dendrogram with nine objects. And there exists the following hierarchical structure H.

$$H = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}$$
$$\{1,2\} \quad \{5,6\} \quad \{5,6,7\} \quad \{3,4\} \quad \{8,9\}$$
$$\{5,6,7,8,9\}, \{1,2,3,4\}, \{1,2,3,4,5,6,7,8,9\}\}$$

A method which transforms a D or S into a hierarchical structure H may be regarded as a procedure which imposes the ultrametric property of dissimilarity or similarity, whether the original is metric or non-metric. In particular, the AHC methods are procedures which form the <H,h> with a monotonic hierarchical structure. We can then obtain a hierarchical partition

$$C^{\alpha} = \{C_1, C_2, \ldots, C_{n-\alpha}\} \quad (\alpha=0,1,2,\ldots,n-1)$$

at the level h_{α}, which is derived from $C^{\alpha-1}$ at the level $h_{\alpha-1}$. The ultrametric is a property of the dissimilarity or similarity which satisfies the following expression as the condition of transitivity:

$$\delta_{ij} \le \max\{\delta_{ik}, \delta_{kj}\} \text{ for any } i,j,k \in E \tag{1}$$

or

$$\delta'_{ij} \ge \min\{\delta'_{ik}, \delta'_{kj}\} \text{ for any } i,j,k \in E \tag{1'}$$

To distinguish between (1) and (1)', the inequality (1)' is sometimes said to be an inframetric. These inequalities may be always derived from the dendrogram. Hence, continuing our discussion, we consider the two well-known methods of single linkage and complete linkage and, of course, the both methods possess the above described property. Both procedures define the distance between two clusters C_p, $C_q \in C^{\alpha-1}$, which are fused at the level h_{α}, as follows:

Single linkage method

$$\min\{d_{ij} \mid O_i \in C_p, O_j \in C_q\} \tag{2}$$
$$\triangleq \min_{l,m} [\min_{r,s}\{d_{rs} \mid O_r \in C_l,\ O_s \in C_m\}] \quad (1 \le l, m \le n-\alpha, l \neq m, p \neq q)$$

Complete linkage method

$$\max\{d_{ij} \mid O_i \in C_p, O_j \in C_q\} \tag{3}$$
$$\triangleq \min_{l,m} [\max_{r,s}\{d_{rs} \mid O_r \in C_l, O_s \in C_m\}] \quad (1 \le l, m \le n-\alpha, l \neq m, p \neq q)$$

where the symbol "$\triangleq$" indicates the definition.

The most essential difference between these two methods is that complete linkage requires maximum operation and single linkage requires a minimum operation. In other words, complete linkage is exactly the opposite of single linkage. Turning our attention to the feature that both methods are characterized only by a maximum or minimum operation, we try to introduce the concepts of fuzzy set theory, in particular fuzzy relation or fuzzy graph into the generalized extension of AHC methods. It is our next aim to examine the relationship between fuzzy relations and AHC methods.

We shall now define the subset A of E to which $\mu(i|A)$ or μ_i represents the degree of belongingness. Under the considerations of ordinary set theory, we can consider that if any $i \in A$ then $\mu_i=1$, and if any $i \notin A$ then $\mu_i=0$, and say, μ_i is a characteristic function. But if the value of μ_i takes in the interval [0,1], μ_i is called a membership function. A subset A of this kind is said to be a fuzzy subset. We assume two fuzzy subsets A,B and define them as follows:

$$A \subseteq B \quad \text{iff} \quad \mu(i|A) \leq \mu(i|B) \quad \text{for any } i \in E$$
$$A \wedge B = \{(i, \min\{\mu(i|A), \mu(i|B)\}) | i \in E\} \qquad (4)$$
$$A \vee B = \{(i, \max\{\mu(i|A), \mu(i|B)\}) | i \in E\}$$

Therefore, the operators $\vee$ and $\wedge$ stand for union and intersection in the sense of fuzzy set theory, that is, $\vee$ and $\wedge$ indicate the maximum and minimum, respectively. And we can define also a fuzzy relation in $E_1 \times E_2$ as follows:

$$R = [\{(i.j), \mu(i,j|R)\} | i \in E_1, j \in E_2] \qquad (5)$$

In particular, if $E_1=E_2=E$, we have the following fuzzy (binary) relation.

$$R = [\{(i,j), \mu(i,j|R)\} | i,j \in E] \qquad (6)$$

where $\mu(i,j|R)$ is a membership function which represents the degree of belongingness of pair (i,j) to the subset $E^2 = E \times E$. We suppose that the value of $\mu(i,j|R)$, in abbreviation $\mu(i,j)$ or μ_{ij}, takes only in the interval [0,1]. Then there are many fuzzy relations with various conditions. We shall define the condition of some fuzzy relations as follows:

(a) $\mu_{ii} = 1$ for any $i \in E$ (reflexivity)

(a)' $\mu_{ii} = 0$ for any $i \in E$ (anti-reflexivity)

(b) $\mu_{ij} = \mu_{ji}$ for any $i,j \in E$ (symmetry)

(c) $\mu_{ij} \geq \max_k [\min \{\mu_{ik}, \mu_{kj}\}]$ (7)

for any $i,j,k \in E$ (max-min transitivity)

(d) $\mu_{ij} \leq \min_k [\max \{\mu_{ik}, \mu_{kj}\}]$

for any $i,j,k \in E$ (min-max transitivity)

Table 1. Summary of some fuzzy relations

relation \ condition	(a)	(a)'	(b)	(c)	(d)
similitude	X		X	X	
dissimilitude		X	X		X
resemblance	X		X		
dissemblance		X	X		

As shown in Table 1, we can consider the several fuzzy relations by the suitable combination of each condition. For example, the relation that satisfies the conditions (a), (b), (c) is a similitude relation. Thus, by the aid of the notations described above, we can easily find that the non-metric dissimilarity is identical to the fuzzy dissemblance relation and that the non-metric similarity is identical to the fuzzy resemblance relation.

Additionally, we can see that an ultrametric inequality for the distance is identical to a min-max transitivity, that is, an ultrametric is a fuzzy dissimilitude relation. Similarly, an inframetric is a fuzzy similitude relation. Occasionally, the former is called the dissimilarity relation, the latter is called the similarity relation, and either construct the equivalence relation.

In other words, the similarity or dissimilarity matrix derived from a dendrogram with a monotonic hierarchical structure is an equivalence relation. In particular we can observe that the result of single linkage is identical to a relation represented on a fuzzy (min-max or max-min) closure obtained by repeating the fuzzy (min-max or max-min) composition defined by the right-hand side of the expression (7)-(c) or (7)-(d). Moreover it is obvious that the expression (1) or (1)' is implied by the condition (7)-(d) or (7)-(c) and it in turn implies the expression.

3. A MEASURE OF DIFFERENCE BETWEEN TWO SIMILARITY MATRICES

As mentioned above, the AHC methods that have been proposed up to the present are considered to be exact methods for forming the fuzzy equivalence relation itself by a kind of successive approximation. We are interested, in the following, in evaluating the difference between two dissimilarity or similarity matrices. We need an index which examines the difference between the original similarity matrix and the matrix derived from a dendrogram. Since the dissimilarity or similarity formed by single and complete linkage is regarded as a fuzzy dissimilitude or similitude relation, as the extention of ordinary symmetric difference, it is natural and valid that we consider the fuzzy symmetric difference as a measure of evaluating and comparing the result of clustering process. And this measure is defined as follows.

Let $S = (s_{ij})$ and $S^* = (s^*_{ij})$ denote the original and derived similarity matrix, respectively, where s_{ij} is normalized in the interval [0,1]. Then, the fuzzy symmetric difference is defined as follows:

$$\rho(S,S^*)=(S\wedge\bar{S}^*)\vee(\bar{S}\wedge S^*) \tag{8}$$

where $\bar{S}$ and $\bar{S}^*$ represent the complement of S and S*, respectively. Though we will mainly describe here the case of similarity, our consideration can be easily extended to a case with dissimilarity measures. And if we let ρ_{ij} denote an element of matrix $\rho(S,S^*)$, we then have the following relationship,

$$\begin{aligned}\rho_{ij} &= (s_{ij}\wedge\bar{s}^*_{ij})\vee(\bar{s}_{ij}\wedge s^*_{ij})\\ &= \frac{1}{2}\{1-|s_{ij}+s^*_{ij}-1|+|s_{ij}-s^*_{ij}|\}\end{aligned}$$

accordingly,

$$\text{if } s_{ij}+s^*_{ij}\le 1, \text{ then } \rho_{ij} = \frac{1}{2}\{s_{ij}+s^*_{ij}+|s_{ij}-s^*_{ij}|\}$$

$$\text{if } s_{ij}+s^*_{ij}>1, \text{ then } \rho_{ij} = 1-\frac{1}{2}\{s_{ij}+s^*_{ij}-|s_{ij}-s^*_{ij}|\}$$

where $\bar{s}_{ij} = 1-s_{ij}$, $\bar{s}^*_{ij} = 1-s^*_{ij}$.

Thus, we can obtain

$$\rho_{ij} = \begin{cases}\max(s_{ij},s^*_{ij}) & \text{if } s_{ij}+s^*_{ij}\le 1\\ 1-\min(s_{ij},s^*_{ij}) & \text{if } s_{ij}+s^*_{ij}>1,\end{cases} \tag{9}$$

where the range of ρ_{ij} is in the interval [0,1].

For convenience and simplicity in our considerations, we investigate a degree of goodness of fit between two relations by a measure of $\rho(S,S^*)$, say $r = \|\rho(S,S^*)\| = \sum_{i<j}\rho_{ij}$. In the sense of fuzzy set theory, a difference between S in itself is not always zero, namely, let it be denoted by $\rho_0(S,S)$, $r_0=\|\rho_0(S,S)\| = \|S\wedge\bar{S}\|$ is not always zero. Moreover, the maximum difference is given by $r'=\|S\vee\bar{S}\|$. Thus, we can obtain the expression $r_0\le r\le r'$. Using these results, we shall propose an index as follows.

We shall denote by r* the index which indicates a fuzzy degree of fitness between two relations.

$$r^* = (r - r_0) / (r' - r_0) \tag{10}$$

Obviously, this expression satisfies the inequality $0 \le r^* \le 1$. Therefore, we can examine the degree of the goodness of fit by the value of r or r*. Our considerations based on the fuzzy symmetric difference are regarded as the generalized extension of absolute deviation or stress measure in the sense of statistics.

EXAMPLE 1.

We shall consider the clustering of the sets of data shown as the scatter diagrams in Figures 2-(a), (b) and (c). Data (a) seems to consist of slightly compact groups. Data (b) has a configuration which is considerably vague in shape. And data (c) seems to have a string-like shape. We can obtain the results shown in Table 2 by applying single linkage and complete linkage to these sets of data. In this case, we shall try to estimate subjectively the similarity s_{ij} among the objects by the visual judgement of figures. The matrix obtained by such a way is immediately considered as a non-metric similarity matrix, that is, a fuzzy resemblance relation. Examination of the results obtained from data (a) shows that complete linkage is better than single linkage. In data (c), the results of the two methods are not very different but the value obtained by single linkage is smaller than that produced from data (a). Moreover, investigating the results obtained by applying single linkage to the three data-sets, we can observe that is better fitted solution is obtained in the case of data (b) or (c). The above results illustrate clearly that complete linkage is superior to single linkage in configurations such as data (a), however, that single linkage is better than complete linkage in the case of data (c) or (b). Finally, observing the behavior of index r*, we can evaluate more quantitatively the validity of a clustering process which has been judged by empirical and subjective interpretability as usual. And investigation of r*'s provides a clue that enables us to obtain reasonable solutions.

4. COMPARING PARTITIONS OBTAINED BY CLUSTERING

Though there are many problems to be faced in using cluster analysis in practical, the most important and difficult are to handle the following situations:

i) examining two dendrograms obtained by applying different clustering algorithms to the same data.

ii) comparing and evaluating two dendrograms based on different sets or the same set of data, and examining partitions generated from those dendrograms.

In short, there always exist the problems of comparison between dendrograms and investigating the partitions formed on dendrograms.

COMPARISON BETWEEN TWO DENDROGRAMS

First we shall examine the two dendrograms obtained by applying district clustering algorithms to the same data set. We now denote two dendrograms by $\langle H_A,h\rangle$, $\langle H_B,t\rangle$, and represent the relations (i.e. similitude relations) given by the both dendrograms by R_A, R_B, respectively. Then it is natural to apply the concept of fuzzy symmetric difference described in the above section to this case. That is, we can investigate the relative difference between two dendrograms by the measure $\rho(R_A,R_B)$. We shall verify the validity of our consideration by simple illustrations.

EXAMPLE 2.

We shall put $E=\{1,2,3,4\}$ and denote two dendrograms by $\langle H_A,h\rangle$, $\langle H_B,t\rangle$, namely,

$$H_A = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,4\}, \{2,3\}, \{1,4,2,3\}\}$$
$$\{h_l\} = \{h_0,h_1,h_2,h_3\} = \{1.0,0.8,0.6,0.4\}$$

and

$$H_B = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1,3\}, \{2,4\}, \{1,3,2,4\}\}$$
$$\{t_m\} = \{t_0,t_1,t_2,t_3\} = \{1.0,0.7,0.5,0.3\}.$$

Table 2.

The r*'s computed for the sets of data in Figure 2.

	data set		
Method	(a)	(b)	(c)
complete linkage	0.0717	0.3570	0.2895
single linkage	0.3779	0.2108	0.2250

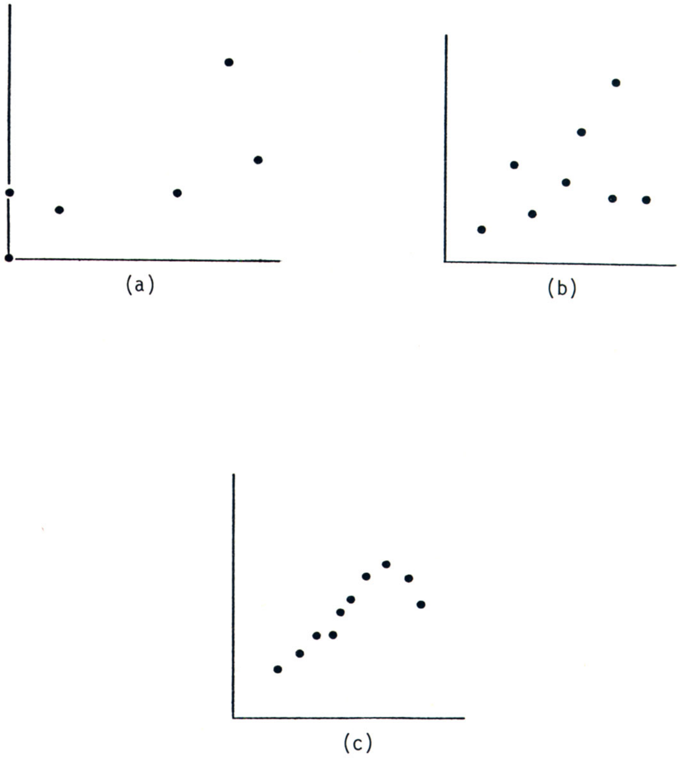


Figure 2. Artificial data

Accordingly, we can obtain two relations, R_A and R_B, from $\langle H_A,h\rangle$ and $\langle H_B,t\rangle$, respectively.

$$R_A = \begin{bmatrix} 1.0 & 0.4 & 0.4 & 0.8 \\ & 1.0 & 0.6 & 0.4 \\ & & 1.0 & 0.4 \\ & & & 1.0 \end{bmatrix} \quad \text{for } \langle H_A,h\rangle$$

$$R_B = \begin{bmatrix} 1.0 & 0.3 & 0.3 & 0.5 \\ & 1.0 & 0.7 & 0.3 \\ & & 1.0 & 0.3 \\ & & & 1.0 \end{bmatrix} \quad \text{for } \langle H_B,t\rangle$$

Therefore, using the expression (8), we can obtain

$$\rho(R_A,R_B)=(\rho_{ij}) = \begin{bmatrix} 0.0 & 0.4 & 0.4 & 0.5 \\ & 0.0 & 0.4 & 0.4 \\ & & 0.0 & 0.4 \\ & & & 0.0 \end{bmatrix}$$

and

$$r = \sum_{i<j} \rho_{ij} = 2.5.$$

Thus we can see the relative difference or association between R_A and R_B.

COMPARING THE PARTITIONS GENERATED FROM THE TWO DENDROGRAMS

Second, we shall think of a procedure which compares the sets of partitions generated from the two dendrograms, which are obtained by applying different methods to the same data. Let us again denote two dendrograms by $\langle H_A,h\rangle$, $\langle H_B,t\rangle$ and represent those relations by R_A, R_B. Then these similitude relations may be decomposed in the following form

$$R_A = \bigvee_l h_l \cdot \underline{R}_A(h_l) \qquad (0 \le h_l \le 1; l=1,2,\ldots,n-1)$$
$$R_B = \bigvee_m t_m \cdot \underline{R}_B(t_m) \qquad (0 \le t_m \le 1; m=1,2,\ldots,n-1) \tag{11}$$

where $\underline{R}$ are equivalence relations in the sense of ordinary set theory, and $h_l\underline{R}_A$ or $t_m\underline{R}_B$ shows that all the elements of the ordinary relation $\underline{R}_A$ or $\underline{R}_B$ are multiplied by h_l or t_m. For example, if

$$R = \begin{bmatrix} 1.0 & 0.3 & 0.2 & 0.5 \\ & 1.0 & 0.2 & 0.3 \\ & & 1.0 & 0.2 \\ & & & 1.0 \end{bmatrix}$$

then,

$$h_0 = 1.0,\ h_1 = 0.5,\ h_2 = 0.3,\ h_3 = 0.2.$$

Thus,

$$R = \bigvee_l h_l \cdot \underline{R}(h_l)$$

$$= \max \left\{ \quad 1.0 \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.0)} , \quad 0.5 \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}_{\underline{R}(0.5)} \right.$$

$$\left.\underset{\underline{R}(0.3)}{0.3 \cdot \begin{bmatrix} 1 & 1 & 0 & 1 \\ & 1 & 0 & 1 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}}, \quad \underset{\underline{R}(0.2)}{0.2 \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}} \right\} \qquad (12)$$

In particular, we try to cut the two dendrograms at a same level $\alpha (0 \leq \alpha \leq 1)$. And we assume $h_l > \alpha > h_{l+1}$, $t_m > \alpha > t_{m+1}$ for the cut at the level α. Then we can obtain two partitioning sets,

$$\left.\begin{aligned} C_A^l &= \{A_1, A_2, \ldots, A_K\} \quad \text{where } K = n-l \\ C_B^m &= \{B_1, B_2, \ldots, B_L\} \quad \text{where } L = n-m \end{aligned}\right. \qquad (13)$$

This situation, especially in the case of K=3 and L=4, may be shown schematically as Figure 3.

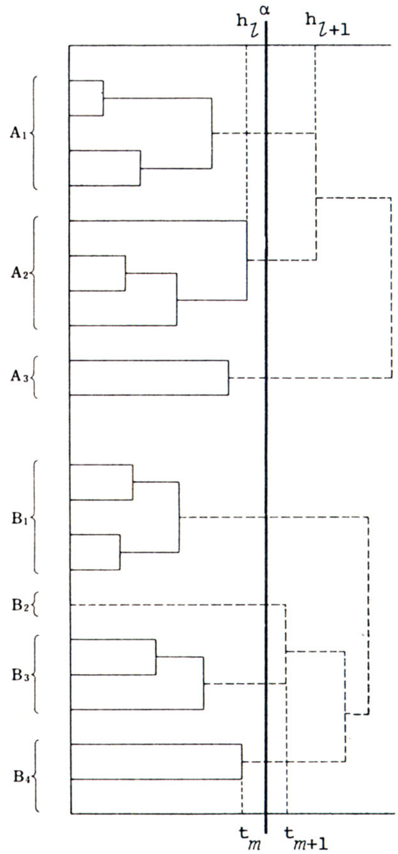


Figure 3. Comparison of the set of partitions produced from the two dendrograms.

Therefore, using the relationship of decomposition for a similitude relation, namely (11), we can generate the two relations

$$R_A^* = \bigvee_l h_l \underline{R}_A(h_l) \qquad (\alpha < h_l)$$
$$R_B^* = \bigvee_m t_m \underline{R}_B(t_m) \qquad (\alpha < t_m) \tag{14}$$

In this case, it is reasonable to consider $\rho(R_A^*, R_B^*)$ as an index for the comparison between the two partitions. However, if we turn our attention to the connectedness between objects rather than the difference between trees, it may be seen that it is natural to use the intersection of two relations, say R_A^* and R_B^* (of course in the sense of fuzzy set theory). Thus the next relationship can be defined,

$$\tau(R_A^*, R_B^*) = R_A^* \wedge R_B^*. \tag{15}$$

Furthermore let τ_{ij} denote an element of matrix $\tau(R_A^*, R_B^*)$ and we can obtain

$$\tau^* = \sum_{i<j} \tau_{ij} \quad . \tag{16}$$

To examine clearly what has been described previously, we shall illustrate with the following example.

EXAMPLE 3.

Let R_A and R_B cite from Example 2 and set the level of cut at $\alpha = 0.45$. Then,

$$R_A^* = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.8 \\ & 1.0 & 0.6 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix}$$

$$R_B^* = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.5 \\ & 1.0 & 0.7 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix}$$

accordingly, by (15),

$$\tau(R_A^*, R_B^*) = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.5 \\ & 1.0 & 0.7 & 0.0 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix} ,$$

and we can obtain $\tau^* = 1.3$.

On the other hand, if we calculate $\rho(R_A^*, R_B^*)$ using (8)

$$\rho(R_A^*, R_B^*) = (\rho_{ij}) = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.5 \\ & 0.0 & 0.4 & 0.0 \\ & & 0.0 & 0.0 \\ & & & 0.0 \end{bmatrix}$$

and $r = \sum_{i<j} \rho_{ij} = 0.9$.

In addition, we shall consider another relation

$$R_c = \begin{bmatrix} 1.0 & 0.3 & 0.5 & 0.3 \\ & 1.0 & 0.3 & 0.7 \\ & & 1.0 & 0.3 \\ & & & 1.0 \end{bmatrix} .$$

Then, $\tau^* = 0$ by using $\tau(R_A^*, R_C^*)$, moreover $r = 2.6$ by using $\rho(R_A^*, R_C^*)$.
In conclusion, we can find that τ^* indicates a kind of degree of agreement between two partitions. That is, if τ^* is large then the construction of the two partitions is similar to each other, if τ^* is small then they may be considered as opposites. In addition, r indicates a deviation or a kind of error between relations formed by the two partitions.

5. EXTENSION OF SINGLE LINKAGE AND COMPLETE LINKAGE

We shall now attempt to modify the algorithm of single and complete linkage, and to extend it to more general case. Let us now define the similarity (or dissimilarity) measures between clusters used by AHC techniques as represented by the following recurrence formula.

$$s_{tr} = \frac{1}{2}(s_{pr} + s_{qr}) + (\gamma - \frac{1}{2})\,|s_{pr} - s_{qr}| \qquad (17)$$

where s_{tr} is the similarity between a cluster C_r and a cluster C_t formed by the fusion of cluster C_p and C_q, and s_{ij} is the similarity between clusters C_i and C_j $(i, j = p,q,r,t)$. γ is a parameter and its value is given beforehand in the interval [0,1]. If $\gamma = 0$, we can obtain complete linkage and if $\gamma = 1$, then single linkage. Moreover, if $\gamma = 1/2$, then the above relation shows the so-called weighted pair group (WPG) method proposed by Sokal. Obviously, all of the results given by applying the above formula to the data, which are dendrograms, have the monotonic hierarchical structure. Therefore, by using the various values of γ, clustering schemes with distinct characteristics can be obtained. If we attempt to adjust the value of γ while keeping the value defined by the expression (10) as small as possible, then we can investigate the solution which is more reasonably fitted to a given data.

Thus, it has been shown that our approach includes a natural generalization and extension for many AHC methods, especially which are similar to single linkage and complete linkage. And we shall call this method modified linkage technique.

EXAMPLE 4.

Now we shall attempt to apply our proposed procedure to Peay's data. This example is from the set of data used by Peay (1975), which in turn is taken from Parkman and Sawyer (1967). The raw data consisted of the numbers of marriages occuring between members of different ethnic groups in Hawaii. The measure is normalized for overall marriage rates which are adjusted to indicate a kind of disparity measure. But in our illustration this measure is transformed into an agreement rate. Accordingly the larger the value, the larger the intergroup marriage rate. The name of ethnic groups included (i.e. objects), and the numbers identified with them are listed as follows:

O_1 : Hawaiian	O_2 : Part-Hawaiian	O_3 : Caucasian
O_4 : Puerto Rican	O_5 : Fillipino	O_6 : Chinese
O_7 : Japanese	O_8 : Korean	

The given raw data is shown in Table 3.

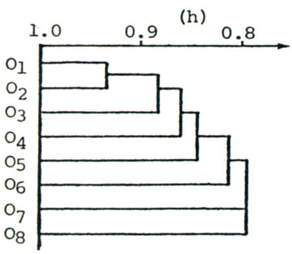
First we shall examine the results obtained by applying single linkage and complete linkage to the similarity matrix in Table 3. And the dendrograms as shown in Figures 4, 5 are produced from single linkage and complete linkage. The fuzzy degree of fitness r^*'s indicated the following values.

i) if complete linkage, $r^* = 0.173$
ii) if WPG method, $r^* = 0.058$
iii) if single linkage, $r^* = 0.000$

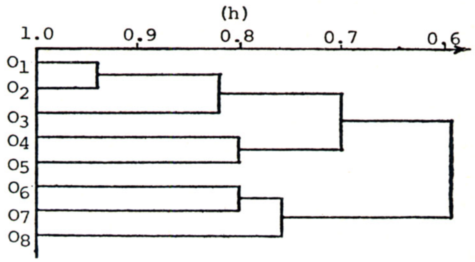
Therefore it is easily seen that there exists a well fairly fitted solution between single linkage and WPG. However we can observe the best fitted solution in the case of single linkage in the sense of fuzzy theory. However observing the dendrogram formed by single linkage, we can investigate the existence of the so-called chaining-effect. Furthermore it also shows no large changes in hierarchical

Table 3. The similarity matrix S for Peay's example.

	1	2	3	4	5	6	7	8
1	1.00	0.94	0.79	0.70	0.82	0.73	0.67	0.68
2		1.00	0.88	0.79	0.86	0.84	0.77	0.77
3			1.00	0.80	0.78	0.76	0.76	0.80
4				1.00	0.81	0.63	0.59	0.63
5					1.00	0.70	0.70	0.72
6						1.00	0.76	0.79
7							1.00	0.80
8								1.00



Dendrogram formed by single linkage
Figure 4.



Dendrogram formed by complete linkage
Figure 5.

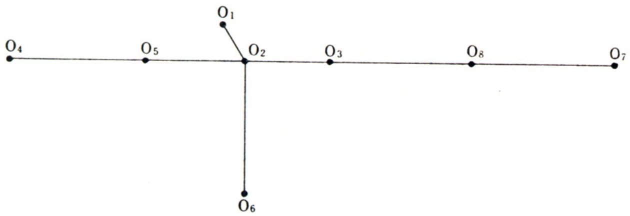


Figure 6. MST generated from the matrix of Table 3.

level. Accordingly, in general, we have determined that such a situation is undesirable because it indicates that the data contains no structure.

On the other hand, as shown in Figure 5, complete linkage produces a dendrogram with large in level, especially going from three groups to two groups. Furthermore we have thought that explicit clusters may exist. However this is doubtful on the basis of additional information obtained from the investigation of the minimum spanning tree (MST) formed from the solution of single linkage. Therefore, in the following, we have constructed a MST. The result is shown in Figure 6. This enables us to examine visually and intuitively relationships between objects.

For example, there exists a slight connectedness between O_4 and O_7, but O_1 and O_2 are very closely related. Furthermore, we can observe a similar situation between O_2 and O_5, or O_2 and O_3. In this example, the link-like information between the objects plays an important role in interpreting and exploring the tendency of data.

In the above discussion, we must pay attention to the property that each value of s_{ij} in the original similarity matrix is larger than 1/2. That is, considering from a fuzzy-theoretic point of view, each object is very similar. Then r* obtained from single linkage always indicates zero. On the contrary, if all values of s_{ij} are smaller than 1/2, then each object is fairly dissimilar. Then r* obtained from complete linkage always indicates zero.

Such properties agree with the well-known features that complete linkage provides a reasonable solution in cases such as the wide-spread or well-separable configurations and that single linkage is superior to complete linkage in configurations such as the agglomerate into a mass or string-like shape.

6. CONCLUSION

In conclusion, we shall attempt to summarize some of the suggestions described already in the previous sections. Above all our main purpose has been to examine several properties which characterize the AHC methods, especially single linkage and complete linkage.

Firstly, the arrangement of AHC methods suggests the fact that many methods have similar features in common. We have discussed consistently the generalized extension of these properties using the fuzzy set theory. Thus, it has been shown that our approach includes a natural generalization and extensions for many AHC methods, especially those which are similar to single linkage and complete linkage.

Next, we proposed that a degree of fitness between solutions by AHC methods and the similarity or dissimilarity of original data is investigated by a fuzzy symmetric difference. And an indicator, say fuzziness r*, derived from a fuzzy symmetric relation makes possible comparisons among the methods. Our consideration is the extension of evaluating procedures based on an ordinary relation, as for example are several works by Jardine and Sibson (1971), Lerman (1970), and Zahn (1969).

Finally we discussed the problems of comparison between dendrograms and investigating the partitions formed on dendrograms, and proposed a practical procedure, in which may be observed the correspondence between the dendrograms (i.e. equivalence relations) and which examines the goodness of fit between partitions generated from dendrograms. This approach is regarded as the generalization of proposals by Rand (1971) and Frank (1977) for comparing partitions. And an examination of several experiments has shown that our proposal is available and useful. Thus we may obtain such reasonable indicators that we have overcome systematically many difficult problems included in most AHC methods, which have been said to be empirical and intuitive up to now.

REFERENCES

[1] Jardine, N. and Sibson, R. (1971): *Mathematical Taxonomy*, John Wiley.

[2] Kaufmann, A. (1973): *Introduction à la Théorie des Sous-ensembles Flous*, Tome I, Masson et Cie.; Tome II (1975), Tome III (1975), Tome IV (1977).

[3] Lerman, I.C. (1970): *Les Bases de la Classification Automatique*, Gauthier-Villars.

[4] Matusita, K. and Ohsumi, N. (1978): Evaluation Procedure of Clustering Techniques, *France-Japan Seminar,* Paris, March, 13th - 20th.
[5] Frank, O. (1976): Comparing Classifications by Use of the Symmetric Class Difference, *COMPSTAT* 1976, pp.89-96.
[6] Rand, W.M. (1971): Objective Criteria for the Evaluation of Clustering Methods, *J. Am. Statist. Ass.,* Vol.66, pp.846-850.
[7] Zadeh, L.A. (1965): Fuzzy sets, *Information and Control,* Vol.8, pp.338-353.
[8] Zadeh, L.A. (1971): Similarity Relations and Fuzzy Orderings, *Inf. Sciences,* Vol.3, pp.177-200.
[9] Zahn, C.T. (1969): Approximating Symmetric Relations by Equivalence Relations, *J. Soc. Ind. Appl. Math.,*Vol.12, No.4, pp.840-847.