A MEASURE OF DISTANCE FOR CLUSTER ANALYSIS BASED ON FUZZY SETS

S. Miyamoto

January 1987

WP-87-42

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

FOREWORD

This paper deals with a measure of distance for cluster analysis. The distance here is defined as a generalization of similarity measures of binary variables using fuzzy set theory. It is proved that the distance introduced in this paper satisfies the triangular inequality. Two algorithms are developed and their convergence is proved.

Alexander B. Kurzhanski
Chairman,
System and Decision Sciences Program

A Measure of Distance for Cluster Analysis Based on Fuzzy Sets

S. Miyamoto

1. Introduction

Since cluster analysis is a convenient and flexible tool of analysis for studying complex systems, it has been used in various fields of systems analysis. When we wish to perform cluster analysis, we must consider carefully the following three points:

1. how to define an appropriate similarity (or distance) measure between a pair of elements to be clustered,
2. which of the algorithms to generate clusters should be selected,
3. how to interprete the generated clusters.

The present paper is concerned mainly with the first point, with the emphasis on a new viewpoint of developing a family of similarity measures based on fuzzy sets.

Up to now, major part of literature in cluster analysis has been concentrated on development of algorithms that is based on the Euclid distance, i.e., the square distance on a finite dimensional space, since we can use various good properties of the Euclid geometry. This means, on the other hand, that multivariate statistical data have been regarded as points in a

multidimensional Euclid space. This assumption is for the purpose of convenience than appropriateness, in order to apply many algorithms that has been developed on the Euclid spaces.

This assumption that an individual data unit should be regarded as a point in the Euclid space is inappropriate, in particular, when the data is binary. Therefore special measures for binary data have been studied (Anderberg, 1973, Everitt, 1980). Frequently attributes of data are described by mixture of variables such as binary, qualitative, and quantitative ones. (cf. Gower, 1971) In some occasions city block distance is prefered than the Euclid distance. (Carmichael and Sneath, 1969) In this way, frequently we should assume a mathematical model that is not based on the Euclid space. In the latter cases of "non-Euclid" models for similarity measures, studies are much more empirical and fewer number of algorithms have been considered.

In a former study (Miyamoto and Nakayama, 1986) we proposed a new type of set-theoretical model to define a family of similarity measures based on the fuzzy set theory. It enables to generalize the similarity measures of binary variables to interval or ratio variables. For example, if we denote by (x_{ik}) the score of the individual i on the variable k, then "fuzzy Jaccard coefficient" s(i,j) between individuals i and j can be defined as

$$s(i,j) = \frac{\Sigma_k \min(x_{ik}, x_{jk})}{\Sigma_k \max(x_{ik}, x_{jk})}$$

Moreover the above consideration leads to a study of a new type of algorithms.

The present paper is concerned with another kind of similarity (or distance) measures based on the framework of fuzzy sets. We begin with the definition of a fuzzy relation defined as a ratio of the intersection and the union of two fuzzy subsets. The similarity measure is introduced as an integral of this fuzzy relation. The new measure of similarity is transformed into a distance measure that is equivalent to the former for the purpose of clustering. The distance measure is a generalization of the Russel and Rao's coefficient or the simple matching coefficient. Moreover the distance measure satisfies the triangular inequality. Two algorithms with their property of convergence are described. These algorithms use two kinds of representative points for a cluster that is called here as centers. It is shown that the two centers can be obtained by simple calculations. It should be noted that the whole argument here suggests, not only the new specific measure of similarity and its related algorithms, but also a new approach to build a mathematical model for clustering based on a set theoretic argument on binary measures. Therefore the present approach implies possibility of developing various kinds of similarity measures of fuzzy relations as future studies.

2. A distance measure based on the framework of fuzzy sets

Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of individuals to be clustered and $Y = \{y_1, y_2, \ldots, y_n\}$ be a set of variables. A

function $h : X \to [0, 1]^Y$ is assumed to be given that maps each individual to its corresponding fuzzy subset in Y as the set of scores that shows relevance of an individual for each variable. Moreover let A and B be two fuzzy subsets in Y. We begin with a definition of a fuzzy relation S(A,B) given as the ratio of $A \cap B$ and $A \cup B$ that shows similarity between A and B. Let

$$f_{S(A,B)}(y) = \begin{cases} f_{A \cap B}(y)/f_{A \cup B}(y), & y \in \mathrm{supp}(A \cup B) \\ 1, & y \notin \mathrm{supp}(A \cup B) \end{cases}$$

be the membership function of S(A,B). Another relation S'(A,B) which is closely related to the former is given by

$$f_{S'(A,B)}(y) = \begin{cases} f_{A \cap B}(y)/f_{A \cup B}(y), & y \in \mathrm{supp}(A \cup B) \\ 0, & y \notin \mathrm{supp}(A \cup B) \end{cases}$$

The difference between S and S' depends on our consideration on a variable y with $f_A(y) = f_B(y) = 0$. In the former definition of S we assume that the value of similarity on this variable should be unity because $f_A(y)$ and $f_B(y)$ take the same value zero. On the other hand in the latter definition we consider that this variable is not interesting and should be ignored.

In the sequel our discussion will be concentrated, not on S and S', but their integrals. Let P be an additive and nonnegative measure such that $\int_Y dP = 1$. We will consider

$$s(A,B) = \int_Y f_{S(A,B)} dP$$

and

$$s'(A,B) = \int_Y f_{S'(A,B)} dP \quad .$$

Remark In a former study (Miyamoto and Nakayama, 1986) we discussed a measure

$$\frac{\int f_{A \cap B}(y) dP}{\int f_{A \cup B}(y) dP}$$

as a generalization of the Jaccard coefficient. On the other hand, the present relations S and S' themselves show similarities between A and B on each variable y. Moreover it will be shown that the integrals s and s' are generalizations of standard binary measures.

It is straightforward to see some properties of s and s'.

Prop.1

$s(A,B) = 1$ iff $A = B$

$s'(A,B) = 1$ iff $A = B = Y$

$s(A,B) = 0$ iff $B = A^C$ and also A and B are crisp

$s'(A,B) = 0$ iff $A \cap B = \emptyset$

$s(A,B) \geq s'(A,B)$ and the equality holds

iff $supp(A \cup B) = Y$.

(Proof) The above relations immediately follow from

$f_{S(A,B)} \equiv 1$ iff $A = B$

$f_{S'(A,B)} \equiv 1$ iff $A = B = Y$

$f_{S(A,B)} \equiv 0$ iff $B = A^C$ and A and B are crisp

$f_{S'(A,B)} \equiv 0$ iff $A \cap B = \emptyset$. []

In the following we write $s(x_i,x_j)$, $s'(x_i,x_j)$ instead of

$s(h(x_i),h(x_j))$, $s'(h(x_i),h(x_j))$ for simplicity.

Now, let us assume that $f_{h(x_i)}(y_k) = x_{ik} \geq 0$, $1 \leq i \leq m$, $1 \leq k \leq n$. If we use the standard sup-min definition for the fuzzy set operation, we have

$$f_{S(x_i,x_j)}(y_k) = \begin{cases} \dfrac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})}, & \max(x_{ik},x_{jk})>0 \\ 1, & x_{ik}=x_{jk}=0 \end{cases}$$

$$f_{S'(x_i,x_j)}(y_k) = \begin{cases} \dfrac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})}, & \max(x_{ik},x_{jk})>0 \\ 0, & x_{ik}=x_{jk}=0 \end{cases}$$

if we define n_{ij}= (number of elements in $\mathrm{supp}(h(x_i) \cup h(x_j))$ and if we assume that $P(y_k) = 1/n$ for all k, then we have

$$s(x_i,x_j) = \frac{1}{n}\left[\sum_k{}' \frac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})} + (n - n_{ij}) \right]$$

$$s'(x_i,x_j) = \frac{1}{n}\left[\sum_k{}' \frac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})} \right]$$

where the summation Σ' is taken over all k such that $\max(x_{ik},x_{jk}) > 0$.

Remark We defined x_i as individuals to be clustered and x_{ik} as a value of membership of $h(x_i)$. Therefore x_i is not a vector whose coordinates are $(x_{i1},\ldots,x_{in})$. The reason why we use this notation which appears somewhat confusing is that in later sections an individual x_i to be clustered can be identified with the vector $(x_{i1},\ldots,x_{in})$ for simplicity and without confusion. []

Prop.2 If we assume that the function $h(x_i)$ is a binary mapping, that is, $x_{ik}= 1$ or $x_{ik}= 0$ for all i and k, then the measure s is equal to the simple matching coefficient (SMC) and the measure s' is equal to the Russel and Rao's coefficient.

Remark We give here the definition of the above two measures of binary variables for the ease of reference. (cf. Anderberg, 1973 or Everitt, 1980) As in the 2 x 2 table below (Table 1), we define the number a as the number of variables that both x_i and x_j have the same value of unity ($x_{ik}=x_{jk}= 1$), b as the number of variables such that $x_i = 1$ and $x_j = 0$, c as the number of variables such that $x_i = 0$ and $x_j = 1$, and d as the number of variables such that $x_i = x_j = 0$.

Table 1

i \ j	1	0
1	a	b
0	c	d

Then the simple matching coefficient is given by (a+d)/(a+b+c+d) and the Russel and Rao's coefficient is given by a/(a+b+c+d).[]

In the sequel the difference between s and s' is not important. Therefore we will consider only the measure s but the argument below can be applied to s' without essential modification. Moreover we use a simplified notation

$$s(x_i,x_j) = \frac{1}{n}\left[\sum_k^n \frac{\min(x_{ik},x_{jk})}{\max(x_{ik},x_{jk})}\right]$$

In the above expression a term $\min(x_{ik},x_{jk})/\max(x_{ik},x_{jk})$ with $\max(x_{ik},x_{jk}) = 0$ should be interpreted as unity. We define a measure of dissimilarity (distance)

$$d(x_i,x_j) = 1 - s(x_i,x_j) .$$

It is immediate to see that $0 \leq d(x_i,x_j) \leq 1$.

Prop.3 The distance measure $d(x_i,x_j)$ satisfies the conditions of metric:

(i) $d(x_i,x_j) \geq 0$, $d(x_i,x_j) = 0$ iff $h(x_i) = h(x_j)$

(ii) $d(x_i,x_j) = d(x_j,x_i)$

(iii) $d(x_i,x_k) + d(x_j,x_k) \geq d(x_i,x_j)$.

(Proof) It is easy to prove the first two properties. For the triangular inequality (iii), note that the following equation holds.

$$d(x_i,x_j) = \frac{1}{n}\sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{\max(x_{ik},x_{jk})}$$

In this equation the term $|x_{ik}-x_{jk}|/\max(x_{ik},x_{jk})$ with $\max(x_{ik},x_{jk}) = 0$ should be considered as zero. Then the triangular inequality follows from the following lemma.

Lemma 1 Let a,b, and c be nonnegative numbers. Then,

$$\frac{|a-c|}{\max(a,c)} + \frac{|b-c|}{\max(b,c)} \geq \frac{|a-b|}{\max(a,b)}$$

where it is assumed that when $\max(a,b) = 0$, the corresponding term should be considered as zero, and so on.

(Proof of Lemma 1) The above inequality can be proved by simple calculations. Let us assume $a \geq b \geq c \geq 0$. Then,

$$\frac{|a-c|}{\max(a,c)} + \frac{|b-c|}{\max(b,c)} - \frac{|a-b|}{\max(a,b)} = \frac{a-c}{a} + \frac{b-c}{b} - \frac{a-b}{a} = \frac{(b-c)(a+b)}{ab} \geq 0$$

In case $a \geq c \geq b \geq 0$, we have

$$\frac{|a-c|}{\max(a,c)} + \frac{|b-c|}{\max(b,c)} - \frac{|a-b|}{\max(a,b)} = \frac{a-c}{a} + \frac{c-b}{c} - \frac{a-b}{a} = \frac{(c-b)(a-c)}{ac} \geq 0$$

If $c \geq a \geq b \geq 0$, then

$$\frac{|a-c|}{\max(a,c)} + \frac{|b-c|}{\max(b,c)} - \frac{|a-b|}{\max(a,b)} = \frac{c-a}{c} + \frac{c-b}{c} - \frac{a-b}{a} = \frac{(c-a)(a+b)}{ac} \geq 0$$

and so on. []

3. Algorithms using centers based on $d(x_i, x_j)$

Although the measure defined in the previous section is directly applicable to standard method of hierarchical agglomerative clustering, we consider in this section some algorithms of nonhierarchical clustering (cf. Anderberg, 1973). Major part of nonhierarchical algorithms use some representative points for clusters, for example, the centroid in case of the Euclid space. Therefore we consider some "centers" of clusters as the representatives of the clusters based on the above $d(x_i, x_j)$.

Let $X_p \subset X$ be a crisp subset (cluster) of X. Without loss of generality we assume that $\{x_1, x_2, \ldots, x_{n_p}\} = X_p$. Let us define z as a vector whose coordinates are $(z_1, \ldots, z_n)$. Since z is not an element of X, it is impossible to consider the distance $d(x_i, z)$. However, as was noted earlier, it will be more convenient from now to identify x_i with a vector $x_i = (x_{i1}, \ldots, x_{in})$ to consider $d(x_i, z)$ for simplicity.

Now, we define a center $\bar{z}^p$ for X_p to be a minimizing element of a problem

$$\begin{aligned} &\min_{z} \sum_{x_i \in X_p} d(x_i, z) \\ &\text{subject to } 0 \le z_k \le 1, \quad k=1,2,\ldots,n \end{aligned} \tag{1}$$

where

$$d(x_i, z) = \frac{1}{n} \sum_{k=1}^{n} \frac{|x_{ik} - z_k|}{\max(x_{ik}, z_k)}$$

We note that the following lemma holds. The latter half of the lemma will be used later.

Lemma 2 Let $0 \le w_1 \le w_2 \le \ldots \le w_n$. Then,

1) the minimizing element $\bar{x}$ of

$$\min_{x \ge 0} J_1(x) = \min_{x \ge 0} \sum_{k=1}^{n} \frac{|w_k - x|}{\max(w_k, x)}$$

is equal to one of w_k's such that

$$J_1(\bar{x}) = \min_{1 \le k \le n} J_1(w_k)$$

2) The minimizing element $\bar{x}$ of

$$\min_{x\geq 0} J_2(x) = \min_{x\geq 0} \max_{w_1,..,w_n} \frac{|w_k - x|}{\max(w_k, x)}$$

is given by $\bar{x} = \sqrt{w_1 w_n}$.

(Proof) Assume $x \geq w_n$, then $J_1(x) = n - \Sigma_k\, w_k/x$.

Therefore $\min_{x\geq w_n} J_1(x) = J_1(w_n)$.

In the same way, $\min_{0\leq x\leq w_1} J_1(x) = J_1(w_1)$

Let $w_i \leq x \leq w_{i+1}$, then

$$J_1(x) = \sum_{k=1}^{i} \frac{x - w_k}{x} + \sum_{k=i+1}^{n} \frac{w_k - x}{w_k}$$

$$= n - \frac{1}{x}\left(\sum_{k=1}^{i} w_k\right) - \left(\sum_{k=i+1}^{n} \frac{1}{w_k}\right) x$$

It is easy to see that the second derivative of $J_1(x)$ is minus.

Hence we have

$$\min_{w_i\leq x\leq w_{i+1}} J_1(x) = \min(J(w_i), J(w_{i+1}))$$

Therefore $\min_{x\geq 0} J_1(x) = \min_{1\leq k\leq n} J_1(w_k)$.

2) Assume $x \geq w_n$, then $J_2(x) = \max_k (x - w_k)/x = (x - w_1)/x$.

Therefore $\min_{x\geq w_n} J_2(x) = J_2(w_n)$.

Assume that $0 \leq x \leq w_1$, then

$J_2(x) = \max_k (w_k - x)/w_k = 1 - x/w_n$, therefore $\min_{0\leq x\leq w_1} J_2(x) = J_2(w_1)$.

Let $w_i \leq x \leq w_{i+1}$, then

$$\max_{w_k} \frac{|w_k - x|}{\max(w_k, x)} = \max\left(\max_{1\leq k\leq i} \frac{x - w_k}{x} , \max_{i+1\leq k\leq n} \frac{w_k - x}{w_k} \right)$$

$$= \max (1 - w_1/x , 1 - x/w_n) .$$

Therefore $\min_{x \geq 0} J_2(x) = \min_{x \geq 0} \max (1 - w_1/x, 1 - x/w_n)$. Now, it is easy to see that $J_2(\sqrt{w_1 w_n}) = \min_{x \geq 0} J_2(x)$

[]

The following proposition is a direct consequence of the above lemma.

Prop.4 Let

$$J_1{}^k(z) = \sum_{i=1}^{n_p} \frac{|z - x_{ik}|}{\max(z, x_{ik})}$$

Then the solution $\bar{z} = (\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n)$ of (1) is given by

$$J_1{}^k(\bar{z}_k) = \min_{x \geq 0} J_1{}^k(x) = \min_{1 \leq i \leq n_p} J_1{}^k(x_{ik}) \qquad (2)$$

$k=1,2,\ldots,n$. []

It is sometimes useful to note the following.

Cor.1 If a group consists of two members, i.e., $X_p = \{x_i, x_j\}$, then the coordinates $\bar{z}_k{}^p$ of $\bar{z}^p$ can take any of the two values x_{ik} and x_{jk}, the coordinates of x_i and x_j, k=1,2,..,n.

Remark The proof of the above corollary is immediate from Prop.4. At the same time it should be noted that a weaker property: $\bar{z}^p = x_i$ or $\bar{z}^p = x_j$ follows from the triangular inequality. []

The solution of (1) in Prop.4 is considered to be a representative point of a cluster. Although this center can be used in agglomerative hierarchical algorithms (cf. Miyamoto and Nakayama, 1986), we show here an algorithm based on an iteration

as a method of nonhierarchical clustering. Since a major part of nonhiearchical clustering deals with optimization problems, we consider here a minimization problem:

$$\min \sum_{p=1}^{Q} \sum_{x_i^p \in X_p} d(x_i^p, \bar{z}^p) \qquad (3)$$

with respect to $X_1, X_2, \ldots, X_Q$ that is a partition of X and $\bar{z}^p$ is the center given by (2) in Prop.4.

The following algorithm is an analogy of Forgy's algorithm on the Euclid space (Anderberg, 1973). It is assumed that the number of clusters are denoted as Q that is given beforehand and is fixed throughout the procedure.

Algorithm 1

1. Begin with initial Q partition.
2. Calculate centers $\bar{z}^p$ for each cluster for all p=1,2,...,Q.
3. Allocate each individual x_i to the cluster with the nearest center. (The nearest center means the center with min $d(x_i, \bar{z}^p)$.)
4. If no individuals change their cluster membership, stop the iteration and output the result. Otherwise go to step 2.

Prop.5 Algorithm 1 converges after a finite number of iteration.
(Proof) It is clear that each iteration of step 2 and step 3 in Algorithm 1 decreases monotonically the value

$$\sum_{p=1}^{Q} \sum_{x_i^p \in X_p} d(x_i^p, \bar{z}^p)$$

therefore the algorithm is convergent after a finite number of

iteration. []

Another consideration on nonhierarchical clustering is based on an idea to deal with each coordinate independently in the calculation of a center. We define another criteria

$$\min_{0\le z_k^p\le 1} \sum_{p=1}^{Q} \sum_{k=1}^{n} \max_{x_i^p \in X_p} \frac{|z_k^p - x_{ik}^p|}{\max(z_k^p, x_{ik}^p)} \qquad (4)$$

$$= \min_{0\le z_k^p\le 1} \sum_{p=1}^{Q} J_3(X_p)$$

where

$$J_3(X_p) = \sum_{k=1}^{n} \max_{x_i^p \in X_p} \frac{|z_k^p - x_{ik}^p|}{\max(z_k^p, x_{ik}^p)}$$

From lemma 2 we have

Prop.6 The solution $(\bar{z}_1^p, \bar{z}_2^p, \ldots, \bar{z}_n^p)$ of (4) is given by

$$\bar{z}_k^p = \sqrt{(\min_{1\le i\le n_p} x_{ik}^p)(\max_{1\le i\le n_p} x_{ik}^p)} \qquad (5)$$

Remark If we consider a geometrical interpretation of the problem (4), a cluster of points might be expressed as the smallest cube which includes all the points inside it and whose center is given by (5). However, it sometimes leads to misunderstanding to consider a geometrical interpretation of this measure, since absolute value (or norm) of x based on the above distance does not have any meaning. It is clear that d(0,x) = 1 for any x with nontrivial coordinates. Since each coordinate can

be dealt with independently, it appears that methods in one-dimensional clustering might be applicable. We must be careful, however, about the above point. For example, the method of histogram can not be applied directly, since the distance of an element from the origin is always unity. []

Thus, we consider another criterion for clustering:

$$\min \sum_{p=1}^{Q} \sum_{k=1}^{n} \max_{x_i^p \in X_p} \frac{|\bar{z}_k^p - x_{ik}^p|}{\max(z_k^p, x_{ik}^p)} \tag{6}$$

with respect to the partition $X_1, X_2, \ldots, X_Q$ of X. We have an algorithm that is based on a consideration of boundary point. A data unit $x_i^p = (x_{i1}^p, \ldots, x_{in}^p)$ is called a boundary of X_p if for some k, $x_{jk}^p = \min_{1 \le j \le n_p} x_{jk}^p$ or $x_{ik}^p = \max_{1 \le j \le n_p} x_{ik}^p$. Therefore we have at most 2nQ boundary points when we consider Q clusters.

Algorithm 2

1. Begin with an initial partition.

2. Calculate $\bar{z}_k^p$, p=1,...,Q, k=1,...,n by (5) that minimizes

$$\max_{x_i^p \in X_p} \frac{|z_k^p - x_{ik}^p|}{\max(z_k^p, x_{ik}^p)}$$

3. For each boundary point x_i^p, if there is another cluster $X_{p'}$ such that

$$J_3(X_p - \{x_i^p\}) + J_3(X_{p'} \cup \{x_i^p\}) < J_3(X_p) + J_3(X_{p'})$$

move x_i^p from the cluster X_p to $X_{p'}$.

4. If no boundary points change their cluster membership, stop the iteration and print the result. Otherwise go to step 2. []

The following proposition is obvious, since algorithm 2 decreases monotonically the criterion (6).

Prop.7 Algorithm 2 converges after a finite number of iteration.

The above algorithm 2 has an advantage that it is sufficient to examine 2nQ boundary points instead of the whole m data units.

4. A simple example

We show a simple example that shows a characteristic of the present measure of distance. Consider three individuals $x_1=(0,1,1)$, $x_2=(1,1,0)$, and $x_3=(4,4,0)$, then it is clear by the definition of the distance that $d(x_1,x_2)=2/3$, $d(x_2,x_3)=1/2$, $d(x_3,x_1)=11/12$. On the other hand if we denote the Euclid distance as d_E we have $d_E(x_1,x_2)=\sqrt{2}$, $d_E(x_2,x_3)=\sqrt{18}$, $d_E(x_3,x_1)=\sqrt{26}$. That is, by the present distance x_2 is nearer to x_3 than to x_1, whereas by the Euclid distance x_2 would be nearer to x_1 than to x_3.

Assume that we wish to divide the above individuals into two clusters, Note that when we consider a group with two elements $\{x_i,x_j\}$, it is clear that the center $\bar{z}$ for this group is given by $\bar{z}=x_i$ or $\bar{z}=x_j$, as we noted earlier. Therefore by the first method based on the present measure of distance we have $X_1=\{x_1\}$, $X_2=\{x_2,x_3\}$,
whereas by the Euclid distance $X_1=\{x_1,x_2\}$, $X_2=\{x_3\}$.

Next, let us consider the second method in the present framework. Let the coordinates of the centers for the groups for the groups $\{x_1,x_2\}$, $\{x_2,x_3\}$, $\{x_3,x_1\}$ calculated by (5) be $\bar{z}_k^{1,2}$,

$\bar{z}_k^{2,3}$, $\bar{z}_k^{3,1}$, respectively. Then from Prop.6 we have $\bar{z}_1^{1,2}= 0$, $\bar{z}_2^{1,2}= 1$, $\bar{z}_3^{1,2}= 0$; $\bar{z}_1^{2,3}= 2$, $\bar{z}_2^{2,3}= 2$, $\bar{z}_3^{2,3}= 0$; $\bar{z}_1^{3,1}= 0$, $\bar{z}_2^{3,1}= 2$, $\bar{z}_3^{3,1}= 0$. Hence it is easy to see that $J_3(\{x_1,x_2\}) = 2$, $J_3(\{x_2,x_3\}) = 1$, $J_3(\{x_3,x_1\}) = 5/2$. Therefore if we divide the individuals into two groups, we have $X_1 =\{x_1\}$, $X_2 =\{x_2,x_3\}$.

Moreover let us consider $x_4=(0,0,1)$ together with x_1, x_2, and x_3, and assume that we wish to divide $\{x_1,x_2,x_3,x_4\}$ into two clusters X_1, X_2. Then by a simple calculation it is straightforward to see that $X_1=\{x_1,x_4\}$, $X_2=\{x_2,x_3\}$ both by the first method and the second method in the present framework, whereas by the Euclid distance we have $X_1=\{x_1,x_2,x_4\}$, $X_2=\{x_3\}$.

5. Conclusion

Many years ago when digital computers were not well developed, major part of statistical tools that were available for multivariate statistical data were based on the Euclid geometry. Therefore until now multivariate statistical data have been considered to be points in a multidimensional Euclid space. With the development of digital computers we have now other type of mathematical tools such as the hierarchical cluster analysis that is more algorithmic than geometrical. These new tools to deal with statistical data implies various possibilities to work with new types of mathematical models. Nevertheless, as noted earlier, major part of the studies on cluster analysis have been devoted to the processing of data in the Euclid space for convenience's sake.

In this paper we have shown another approach of fuzzy set-

theoretic mathematical model. We discussed several properties of the measures derived from the model and also showed algorithms based on the measures. The present approach of fuzzy set framework together with our former result on the fuzzy Jaccard coefficient (Miyamoto and Nakayama, 1986) is applicable to other area of studies than the cluster analysis. For example, we can suggest the use of these measures to multidimensional scaling and information retrieval.

References

Anderberg, M. R. (1973) Cluster analysis for Applications, Academic press.

Carmichael, J. W. and Sneath, P. H. A. (1969) Taxonometric Maps, Syst. Zool., 18, 402-415.

Everitt, B. (1980) Cluster Analysis, 2nd ed., Halsted Press.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties, Biometrics, 27, 857-872.

Miyamoto, S. and Nakayama, K. (1986) Similarity measures based on a fuzzy set model and application to hierarchical clustering, IEEE Trans., Syst., Man, and Cybern., SMC-16, 3, 479-482.