A FUZZY RELATION APPROACH TO SINGLE LINKAGE

MAMAN A. DJAUHARI *

Abstract. *Single linkage is equivalent to sub-dominant ultrametric. Many algorithms are available for constructing these two objects. But all of them, except one which was proposed by Gondran, are very tedious because of the lack of algebraic structure. Gondran used a special algebraic system as theoretical bases. But it seems rather artificial. In this paper, we propose a more formal approach based on fuzzy relation. The main result presented here is the equivalence between sub-dominant ultrametric and the min-max transitive closure of a symmetric and anti reflexive fuzzy relation. This property enables us to construct an easy and efficient algorithm. At the end of this paper we will find its relationship with Gondran's approach.*

Keywords. *Single linkage, sub-dominant ultrametric and Gondran's algebraic system.*

1. Introduction. Indexed hierarchical clusters (IHC) on a set I, with card(I) = n and a dissimilarity d on I, is constructed based on the fact that there is a one to one correspondence between the set of all IHCs on I and the set of all ultrametrics on I (see [5], [8]). Single linkage is one of the most popular method for constructing an IHC on I (see

* Department of Mathematics, Institut Teknologi Bandung, Bandung 40132, Indonesia

[1], [2], [3], [6], [8]). In this method, the distance ∂ between two clusters A and B is defined by

$$\partial (A,B) = \min_{\substack{i \text{ in } A \\ j \text{ in } B}} d(i,j)$$

We find the important role of single linkage in many applications such as in statistics (especially in cluster analysis and in detecting influencial subsets), numerical taxonomy and automatic classification. We know that the algorithms of single linkage are tedious if n is sufficiently large.

Fortunately single linkage is equivalent to sub-dominant ultrametric (see [2]) where, by definition, the sub-dominant ultrametric (USD) u of d is

$$u = \text{Sup } \{\delta \mid \delta \text{ ultrametric} \leq d\}.$$

These facts motivate us to study the USD by means of fuzzy relation approach. In Section 2 we review some important theorems cited in [4]. The main result of this work is presented in Section 3 and Section 4; here it will be shown that the USD is equivalent to the min-max transitive closure . This property enables us to construct an easy and efficient algorithm. At the end of this paper we will find its relationship with Gondran's approach.

2. Transitive closure. All symbols and properties of fuzzy relation we use here are borrowed from [4] and [5]. Suppose $\underset{\sim}{R}$ is a max-min transitive fuzzy relation on I. In other words,

$$\mu_{\underset{\sim}{R}}(x,z) \leq \bigvee_y \{\mu_{\underset{\sim}{R}}(x,y) \wedge \mu_{\underset{\sim}{R}}(y,z)\}$$

for all x, y, z in I.

Max-min transitivity of a fuzzy relation $\underset{\sim}{R}$ can be verified through the following notion of max-min composition o of $\underset{\sim}{R}$.

$\underset{\sim}{R}^{o2} = \underset{\sim}{R} \text{ o } \underset{\sim}{R}$ is a fuzzy relation, where

$$\mu_{\underset{\sim}{R}^{02}}(x,z) = \bigvee_{y} \{\mu_{\underset{\sim}{R}}(x,y) \wedge \mu_{\underset{\sim}{R}}(y,z)\}$$

for all x, y, z in I.

In [4] we have the following properties for checking the max-min transitivity of $\underset{\sim}{R}$.

Theorem 1. *If* $\underset{\sim}{R}^{02} = \underset{\sim}{R}$, *then* $\underset{\sim}{R}$ *is max-min transitive.*

Theorem 2. $\underset{\sim}{R}$ *is max-min transitive if and only if* $\underset{\sim}{R}^{02} \subseteq \underset{\sim}{R}$.

Since I is finite, the max-min transitive closure $\hat{\underset{\sim}{R}}$ of $\underset{\sim}{R}$ has the following representation :

$$\hat{\underset{\sim}{R}} = \underset{\sim}{R} \cup \underset{\sim}{R}^{02} \cup \underset{\sim}{R}^{03} \cup \ldots \cup \underset{\sim}{R}^{ok},$$

for an integer k; $1 \leq k \leq n$, where $\underset{\sim}{R}^{ok} = \underset{\sim}{R} \circ \underset{\sim}{R} \circ \ldots \circ \underset{\sim}{R}$,

k times max-min composition of $\underset{\sim}{R}$.

Those theorems give us the following theorem (see [4]) which is a standard approach to construct the max-min transitive closure.

Theorem 3. *Suppose* $\underset{\sim}{R}$ *is a fuzzy relation on* I. *Let*

$$l^{*}(x,y) = \bigvee_{c \; di\zeta} l(c), \quad \textit{where}$$

a). $\zeta = \{c \mid c = (x = x_{i_1}, x_{i_2}, \ldots, x_{i_r} = y)\}$ *is a chain from* x *to* y*}*

b). $l(c) = \mu_{\underset{\sim}{R}}(x_{i_1}, x_{i_2}) \wedge \mu_{\underset{\sim}{R}}(x_{i_2}, x_{i_3}) \wedge \ldots \wedge \mu_{\underset{\sim}{R}}(x_{i_{r-1}}, x_{i_r})$.

Then $\mu_{\hat{\underset{\sim}{R}}}(x,y) = l^{*}(x,y)$, *for all* x *and* y *in* I.

In practice, this theorem is still difficult to be

implemented. An easier way, for which $\underset{\sim}{R}$ is a dissimilarity, will be shown in the next section.

Like the max-min transitive closure $\hat{\underset{\sim}{R}}$, the min-max transitive closure $\overset{v}{\underset{\sim}{R}}$ of $\underset{\sim}{R}$ has the following representation.

$$\overset{v}{\underset{\sim}{R}} = \underset{\sim}{R} \cap \underset{\sim}{R}^{*2} \cap \underset{\sim}{R}^{*3} \cap \ldots \cap \underset{\sim}{R}^{*k}$$

for an integer k, $1 \leq k \leq n$, where $\underset{\sim}{R}^{*k} = \underset{\sim}{R} * \underset{\sim}{R} * \ldots * \underset{\sim}{R}$, k times min-max composition * of $\underset{\sim}{R}$ and

$$\mu_{\underset{\sim}{R}^{*2}}(x,z) = \underset{y}{\wedge} \{ \mu_{\underset{\sim}{R}}(x,y) \vee \mu_{\underset{\sim}{R}}(y,z) \}$$

The following theorem enables us to work either with $\hat{R}$ or with $\overset{v}{R}$.

Theorem 4. $\overline{\hat{\underset{\sim}{R}}} = \overset{v}{\underset{\sim}{\overline{R}}}$ *and* $\overset{v}{\underset{\sim}{R}} = \hat{\underset{\sim}{\overline{R}}}$.

Corollary. By De Morgan's rule, we have $\hat{\underset{\sim}{R}} = \overline{\overset{v}{\underset{\sim}{\overline{R}}}}$ and $\overset{v}{\underset{\sim}{R}} = \overline{\hat{\underset{\sim}{\overline{R}}}}$.

3. Equivalence between USD and min-max transitive closure. Suppose D is a dissimilarity matrix associated with dissimilarity index d on I. It means that the (i,j)th element of D is d(i,j), where i and j in I. Suppose also that U is an ultrametric matrix on I. Then the (i,j)th element of U is u(i,j), where u.is an ultrametric on I. From fuzzy relation's point of view, we know that

(i) D is a symmetric and anti-reflexive fuzzy relation,

where

$$\mu_D(x,y) = d(x,y) \text{ for all } x \text{ and } y \text{ in } I.$$

(ii) U is a symmetric, anti-reflexive and min-max transitive fuzzy relation. It means that

$$\mu_U(x,z) \leq \bigwedge_y \{\mu_U(x,y) \vee \mu_U(y,z)\}$$

for all x, y and z in I, where $\mu_U(x,y) = U(x,y)$.

In the next paragraph, we propose two propositions that enable us to find the USD more easily. These two propositions are the main result of this work.

Proposition 1. *If* $\underset{\sim}{R}$ *is a dissimilarity on* I, *then* $\overset{\vee}{\underset{\sim}{R}} = \underset{\sim}{R}^{*k}$ *for an integer* k; $1 \leq k \leq n$.

Proof.

From **Theorem 4** we know that $\overset{\vee}{\underset{\sim}{R}} = \overset{\wedge}{\underset{\sim}{R}}$. Hence,

$$\overset{\wedge}{\underset{\sim}{R}} = \overline{(\underset{\sim}{R} \cup \underset{\sim}{R}^{o2} \cup \underset{\sim}{R}^{o3} \cup \ldots \cup \underset{\sim}{R}^{ok})} \text{ for an integer k; } 1 \leq k \leq n.$$

$$= (\overline{\underset{\sim}{R}} \cap \overline{\underset{\sim}{R}^{o2}} \cap \overline{\underset{\sim}{R}^{o3}} \cap \ldots \cap \overline{\underset{\sim}{R}^{ok}}) \text{ by De Morgan's rule.}$$

We know that

$$\mu_{\overline{\underset{\sim}{R}^{o2}}}(x,y) = \max_{(a,b)} \{\mu_{\underset{\sim}{R}^{o2}}(a,b)\} - \mu_{\underset{\sim}{R}^{o2}}(x,y)$$

$$= \alpha - \max_z \{\min \{\mu_{\underset{\sim}{R}}(x,z), \mu_{\underset{\sim}{R}}(z,y)\}\}$$

where $\alpha = \max_{(a,b)} \{\mu_{\underset{\sim}{R}^{o2}}(a,b)\}$

$$= \mu_{\underset{\sim}{R}}(x,x)$$

$$= \alpha - \max_z \{\alpha - \max \{\alpha - \mu(x,z), \alpha - \mu(z,y)\}\}$$

$$= \alpha - \max_{z} \{\alpha - \max \{\mu_{\bar{\underset{\sim}{R}}}(x,z),\ \mu_{\bar{\underset{\sim}{R}}}(z,y)\}\}$$

$$= \min_{z} \{\max \{\mu_{\bar{\underset{\sim}{R}}}(x,z),\ \mu_{\bar{\underset{\sim}{R}}}(z,y)\}\}$$

$$= \mu_{\bar{\underset{\sim}{R}}^{*2}}(x,y)$$

Hence $\overline{R^{o2}} = \bar{R}^{*2}$.

In general we have $\overline{R^{om}} = \bar{R}^{*m}$; m = 1 , 2 , ... , k. Hence,

$$\overset{\vee}{\underset{\sim}{R}} = \underset{\sim}{\bar{R}} \cap \underset{\sim}{\bar{R}}^{*2} \cap \ldots \cap \underset{\sim}{\bar{R}}^{*k}$$

for an integer k; $1 \leq k \leq n$.

Now we show that the right hand side equals $\underset{\sim}{\bar{R}}^{*k}$.

By definition,

$$\mu_{\bar{\underset{\sim}{R}}^{*2}}(x,z) = \bigwedge_{y} \{\mu_{\bar{\underset{\sim}{R}}}(x,y) \vee \mu_{\bar{\underset{\sim}{R}}}(y,z)\},$$

for all x, y and z in I. Especially if y = z, then

$$\mu_{\bar{\underset{\sim}{R}}^{*2}}(x,z) \leq \mu_{\bar{\underset{\sim}{R}}}(x,z) \vee \mu_{\bar{\underset{\sim}{R}}}(z,z)$$

But $\mu_{\bar{\underset{\sim}{R}}}(z,z) = 0$, because $\bar{\underset{\sim}{R}}$ is a dissimilarity. Hence,

$$\mu_{\bar{\underset{\sim}{R}}^{*2}}(x,z) \leq^{\bullet} \mu_{\bar{\underset{\sim}{R}}}(x,z)$$

for all x and z in I or $\underset{\sim}{\bar{R}}^{*2} \subseteq \underset{\sim}{\bar{R}}$.

In general we have $\underset{\sim}{\bar{R}}^{*k} \subseteq \ldots \subseteq \underset{\sim}{\bar{R}}^{*3} \subseteq \underset{\sim}{\bar{R}}^{*2} \subseteq \underset{\sim}{\bar{R}}$.

It implies that $\overset{\vee}{\underset{\sim}{R}} = \underset{\sim}{R}^{*k}$ and the proof is complete.

Corollary. $\overset{\vee}{\underset{\sim}{R}} = \underset{\sim}{R}^{*k} \leq \underset{\sim}{R}$ or $\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y) \leq \mu_{\underset{\sim}{R}}(x,y)$

for all x and y in I.

The second proposition will show us how to construct easily the USD or to realize easily the single linkage.

Proposition 2. *If* $\underset{\sim}{R}$ *is a dissimilarity on* I, *then* $\overset{\vee}{\underset{\sim}{R}}$ *is the USD of* $\underset{\sim}{R}$.

Proof.

Theorem 3 tells us that

$$\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y) = \mu_{\overset{\wedge}{\underset{\sim}{R}}}(x,y) = \max_{c \text{ di } \mathcal{C}} l(c)$$

where $c = (x = x_{i_1}, x_{i_2}, \ldots, x_{i_r} = y)$ is a chain from x to y.

If $\alpha = \mu_{\underset{\sim}{R}}(x,x)$ for all x in I, then

$$\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y) = \max_{c} \{\min_{k} \{\mu_{\underset{\sim}{R}}(x_{i_k}, x_{i_{k+1}})\}\}$$

$$= \max_{c} \{\min \{\mu_{\underset{\sim}{R}}(x_{i_1}, x_{i_2}), \ldots, \mu_{\underset{\sim}{R}}(x_{i_{r-1}}, x_{i_r})\}\}$$

$$= \max_{c} \{\alpha\text{-max} \{\alpha\text{-}\mu_{\underset{\sim}{R}}(x_{i_1}, x_{i_2}), \ldots, \alpha\text{-}\mu_{\underset{\sim}{R}}(x_{i_{r-1}}, x_{i_r})\}\}$$

$$= \max_{c} \{\alpha\text{-max } \{\mu_{\underset{\sim}{R}}(x_{i_1}, x_{i_2}), \ \ldots \ , \ \mu_{\underset{\sim}{R}}(x_{i_{r-1}}, x_{i_r})\}\}$$

$$= \alpha\text{-}\min_{c} \{\alpha\text{-}\{\alpha\text{-max } \{\mu_{\underset{\sim}{R}}(x_{i_1}, x_{i_2}), \ \ldots \ , \ \mu_{\underset{\sim}{R}}(x_{i_{r-1}}, x_{i_r})\}\}\}$$

$$= \alpha\text{-}\min_{c} \{\max_{k} \{\mu_{\underset{\sim}{R}}(x_{i_k}, x_{i_{k+1}})\}\}$$

This equality shows that

$$\mu_{\underset{\sim}{\bar{R}}^{V}}(x,y) = \min_{c} \{\max_{k} \{\mu_{\underset{\sim}{\bar{R}}}(x_{i_k}, x_{i_{k+1}})\}\}$$

Now we will show that $\underset{\sim}{\bar{R}}^{V}$ is the USD of $\underset{\sim}{\bar{R}}$.

(i). It is clear that $\mu_{\underset{\sim}{\bar{R}}^{V}}(x,y) \leq \mu_{\underset{\sim}{\bar{R}}}(x,y)$ for all x and y in I, since $\underset{\sim}{\bar{R}}^{V} = \underset{\sim}{\bar{R}}^{*k} \subseteq \underset{\sim}{\bar{R}}$ (see **Corollary** to **Proposition 1**).

(ii). If $c = (x = x_{i_1}, x_{i_2}, \ldots, x_{i_r} = y)$ is a chain from x to y, we note that $L(c) = \max_{k} \{\mu_{\underset{\sim}{\bar{R}}}(x_{i_k}, x_{i_{k+1}})\}$.

Suppose c_1 is a chain from x to y and c_2 is a chain from y to z, such that $\mu_{\bar{R}^{V}}(x,y) = L(c_1)$ and $\mu_{\underset{\sim}{\bar{R}}^{V}}(y,z) = L(c_2)$.

Suppose also that c_3 is a chain from x to z, constructed from c_1 and c_2 such that

$$L(c_3) = \max \{L(c_1), L(c_2)\}$$

In this case,

$$L(c_3) = \max \{\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y),\ \mu_{\overset{\vee}{\underset{\sim}{R}}}(y,z)\}$$

Then we have

$$\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,z) = \min_{c\ di\zeta} L(c) \le L(c_3).$$

$$\le \max \{\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y),\ \mu_{\overset{\vee}{\underset{\sim}{R}}}(y,z)\}$$

It implies that $\overset{\vee}{\underset{\sim}{R}}$ is an ultrametric on I.

(iii). Suppose U is the USD of $\overset{-}{\underset{\sim}{R}}$. Now we show that $U = \overset{\vee}{\underset{\sim}{R}}$. Consider a chain $c_1 = (x = x_{i_1}, x_{i_2}, \ldots, x_{i_r} = y)$ from x to y where $\mu_{\overset{\vee}{\underset{\sim}{R}}}(x,y) = L(c_1)$. Then,

a). $\mu_U(x,y) \le \max \{\mu_U(x,z),\ \mu_U(y,z)\}$ for all x, y and z in I, because U is an ultrametric. Especially,

$$\mu_U(x,y) \le \max \{\mu_U(x, x_{i_k}),\ \mu_U(x_{i_k}; y)\}$$

for all k = 1, 2, ..., r. Hence,

$$\mu_U(x,y) \le \max \{\mu_U(x, x_{i_2}),\ \mu_U(x_{i_2}; y)\}$$

$$\le \max \{\mu_U(x, x_{i_2}),\ \max \{\mu_U(x_{i_2}, x_{i_3}),\ \mu_U(x_{i_3}, y)\}\}$$

$$\le \max \{\mu_U(x, x_{i_2}),\ \mu_U(x_{i_2}, x_{i_3}),\ \mu_U(x_{i_3}, y)\}$$

In general we have

$$\mu_U(x,y) \le \max_k \{\mu_U(x_{i_k}, x_{i_{k+1}})\},\ 1 \le k \le r-1.$$

b). U is the USD of $\underset{\sim}{\bar{R}}$. Then by definition, $U \subseteq \underset{\sim}{\bar{R}}$ or

$$\mu_U(x,y) \le \mu_{\underset{\sim}{\bar{R}}}(x,y), \text{ for all } x \text{ and } y \text{ in } I.$$

From a) and b), we have

$$\mu_U(x,y) \le \max_k \{\mu_{\underset{\sim}{\bar{R}}}(x_{i_k}, x_{i_{k+1}})\},\ 1 \le k \le r-1$$

$$\le L(c_1) = \mu_{\underset{\sim}{\bar{R}}^V}(x,y) \text{ or}$$

$$\mu_U(x,y) \le \mu_{\underset{\sim}{\bar{R}}^V}(x,y).$$

It has been shown that $\underset{\sim}{\bar{R}}^V$ is an ultrametric and U is the USD of $\underset{\sim}{\bar{R}}$. Hence the inequality $\mu_u(x,y) \le \mu_{\underset{\sim}{\bar{R}}^V}(x,y)$ gives us

$\mu_u(x,y) = \mu_{\underset{\sim}{\bar{R}}^V}(x,y)$ or $U = \underset{\sim}{\bar{R}}^V$. It implies that $\underset{\sim}{\bar{R}}^V$ is the USD of $\underset{\sim}{\bar{R}}$ and the proof is complete.

4. Conclusion. Proposition 2 shows us how to construct easily the single linkage. Its simple algorithm is given in **Proposition 1**. In practice, if D is a dissimilarity matrix, then the single linkage on D can be easily found by constructing the sequence D^2, D^4, D^8, ... etc, where matrix multiplication is defined as usual but the sum and the product of two real numbers a and b are defined by

$a + b = \min\{a,b\}$ and $a.b = \max\{a,b\}$.

If $D^{2^{(k+1)}} = D^{2^k}$ for an integer k, then D^{2^k} is the result for single linkage on D.

5. Relationship with Gondran's approach. Consider the algebraic system $(R^+, +, .)$ where

$a + b = \min \{a,b\}$ and $a.b = \max \{a,b\}$

for all a and b in R^+. If $\mathcal{S}$ is the set of all nxn matrices over $(R^+, +, .)$, the sum $\oplus$ and the product * of two matrices are defined as usual, we know that $(\mathcal{S}, \oplus, *)$ is the Gondran's algebraic system (see [3], [8], and [9]). By means of the two propositions we find that this algebraic system is the system of fuzzy relation with the two operations V and Λ. Furthermore the sequence D^2, D^4, D^8, ... etc., and thus the solution D^{2^k} where $D^{2^{(k+1)}} = D^{2^k}$ for an integer k, is equal to the solution for single linkage on D given by Gondran (see again [3], [8], and [9]).

REFERENCES

1. Benzecri J.P. *L'analyse des données; la taxinomie.* Dunod - Paris, 1980.
2. Caillez F. and Pages J.P. *Introduction à l'analyse des données.* Smash - Paris, 1976.
3. Jambu M. *Classification automatique pour l'analyse des données.* Dunod - Paris 1978.
4. Kaufmann A. *Introduction à la théorie des sousensemble flous; element théoriques de base.* 2eme edition. Masson Paris 1977.
5. Kaufmann A. *Introduction à la théorie des sousensemble flous; application à la classification et à la reconnaissance des formes aux automates et aux system; aux choix des critères.* Masson - Paris 1975.
6. Jardine N. and Sibson P. *Mathematical taxonomy.* John Wiley and Sons 1977.
7. Pal S.K. and Majumder D.K.D. *Fuzzy Mathematical approach to pattern recognition.* Wiley Eastern Ltd 1986.
8. Roux. *Classification automatique.* Ecole d' Ete d'Analyse Numérique, Paris 1975.

9. Van Cutsem B. *Ultrametrique, distance, ø - distances maximum dominées par une dissimilarité donnée.* Statistique et Analyse des Données, Vol 8 No. 2, 1983.