International Journal of Innovative
Computing, Information and Control
Volume **8**, Number **8**, August **2012**

 ISSN 1349-4198
pp. 5715–5724

FUZZY CLUSTERING ANALYSIS OF DATA MINING: APPLICATION TO AN ACCIDENT MINING SYSTEM

Jianxiong Yang and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135, Japan
leoworldplus@yahoo.co.jp; junzow@osb.att.ne.jp

Received March 2011; revised July 2011

Abstract. *This paper is concerned with the application of data transforms and fuzzy clustering to extract useful data. It is possible to distinguish similar information which includes selector and removes clusters of less importance with respect to describing the data. Clustering takes place in the product space of systems inputs and outputs and each cluster corresponds to a fuzzy IF-THEN rule. By initializing the clustering with a number of clusters and subsequently removing less important ones as the clustering progresses, it is sought to obtain a suitable partition of the data in an automated manner. The approach is generally applicable to the fuzzy-means and related algorithms. In this paper, this method can better return appropriate information for user queries; in particular, a novel ranking strategy is provided to measure the relevance score of an annotated set of web results by considering user queries, data annotation, and the underlying ontology.*
Keywords: Fuzzy clustering, Fuzzy systems, Data mining, Identification

1. **Introduction.** Fuzzy relational equations play important roles in many applications, such as intelligence technology [1]. Therefore, how to compute the solutions of fuzzy relational equations is a fundamental problem. Recently, there have been many research papers investigating the solvability of fuzzy relational equations, by generalizing and extending the original results in various directions [2,3]. In addition, many people have studied the optimization problems with fuzzy relational equation constraints [4,5].

In this paper, we prove that embedding target data into a data set can effectively extract the user's target data using Semantic Web search engines. This sort of ranking exploits the evaluation of accurate information on a Web page. It can be used in conjunction with other ranking strategies to further improve the accuracy of retrieved results. Comparing with other ranking methods for Semantic Web, our approach only depends on the user query, the ranked web pages and the underlying ontology. Thus, it allows us to effectively manage the search space and reduce the complexity of the ranking task.

In this paper, we use the data standardization and max-min tree to develop a new algorithm for extracting useful information. In Section 2 and Section 3 some basic definitions and preliminary theorems are presented. In Section 4, a sufficient and necessary process, for describing intelligent clustering system, is shown. In Section 5, we present an example of algorithm for proving this method. If the given solution is maximal, we hence obtain all minimal solutions. Finally, some concluding remarks are given in Section 6.

2. **Fuzzy Set.** A fuzzy set is a pair (A, m) where A is a set and $m : A \rightarrow [0.\ 1]$. For each $x \in A$, $m(x)$ is called the grade of membership of x in (A, m). For a finite set $A = \{x_1, x_2, \ldots, x_n\}$, the fuzzy set (A, m) is often denoted by $\{m(x_1)/x_1, m(x_2)/x_2, \ldots, m(x_n)/x_n\}$.

Let $x \in A$. Then x is called not included in the fuzzy set (A, m) if $m(x) = 0$, x is called fully included if $m(x) = 1$, and x is called fuzzy member if $0 < m(x) < 1$[]. The set $\{x \in A | m(x) > 0\}$ is called the support of (A, m) and the set is called its kernel.

Sometimes, more general variants of the notion of fuzzy set are used, with membership functions taking values in a (fixed or variable) algebra or structure $\boldsymbol{L}$ of a given kind. The usual membership functions with values in $[0, 1]$ are then called $[0, 1]$-valued membership functions [6].

3. **Fuzzy Clustering.** Fuzzy clustering is a class of algorithm for cluster analysis in which the allocation of data points to clusters. It is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. One of the most widely used fuzzy clustering algorithm is the Fuzzy max-min cluster.

4. **Creating Intelligent Clustering System.**

TABLE 1. Sample data

No.	**Attribute**		
	Economic loss	**Disaster area**	**Casualty state**
	X_1(mil.\$)	X_2(m^2)	X_3
1	13	35	0
2	113	345	46
3	1213	1134	9
4	334	645	5
5	115	235	5
6	288	342	54
7	67	1687	1
8	236	610	6
9	1567	55	62
10	167	567	39
11	334	20	56
12	178	914	11
13	768	44	7
14	213	2	6
15	371	43	85
16	127	232	5
17	74	87	36
18	1123	25	120

TABLE 2. New data set

No.	Attribute		
	Economic loss	**Disaster area**	**Casualty state**
	X_1(mil.\$)	X_2(m^2)	X_3
1	74	87	36
2	1123	25	120
3	13	135	0
4	113	345	46
5	1213	1134	9
6	334	645	5
7	115	235	5
8	288	342	54
9	67	1687	1
10	236	610	6
11	1567	55	62
12	167	567	39
13	334	20	56
14	178	914	11
15	768	44	7
16	213	2	6
17	371	43	85
18	127	232	5
19	200	600	50
20	100	300	5

4.1. **Standardization.**

4.1.1. *Data matrix.* In the data records of data storage, it can establish the classification data set, and quantitative target sets are analyzed. Let $\boldsymbol{M}$ be data set Matrix, where $\boldsymbol{m}_{ij}$ element which is in $\boldsymbol{k}\times\boldsymbol{n}$ array matrix $\boldsymbol{M}$, $\boldsymbol{j} = 1, 2, 3, \ldots, k$; $\boldsymbol{i} = 1, 2, 3, \ldots, n$.

4.1.2. *Data standardization.* In the actual data, usually a different data has different dimensions; therefore, we need to deal with the original data standardization.

a) Translational / standard deviation transformation

$$m'_{ij} = \frac{m_{ij} - \overline{m_j}}{S_j} \tag{1}$$

where $\boldsymbol{i}$ is the number of column, $\boldsymbol{j}$ is the number of row, average $\overline{m_j} = \frac{m_{1j}+m_{2j}+m_{3j}+\cdots+m_{kj}}{k}$, standard deviation $S_j = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(m_{ij} - \overline{m_j})^2}$.

b) Translational / range transformation

After standard deviation transformation, the $\boldsymbol{m}'_{ij}$ is uncertain in the interval [0, 1]. So it requires range transformation.

$$m''_{ij} = \frac{m'_{ij} - \min\limits_{1\le i\le n}\{m'_{ij}\}}{\max\limits_{1\le i\le n}\{m'_{ij}\} - \min\limits_{1\le i\le n}\{m'_{ij}\}} \tag{2}$$

TABLE 3. Standard deviation transformation

No.	Attribute		
	Economic loss	**Disaster area**	**Casualty state**
	X_1(mil.\$)	X_2(m^2)	X_3
1	–0.7035	–0.7129	0.1689
2	1.7078	–0.8537	2.7019
3	–0.8437	–0.6040	–0.9167
4	–0.6139	–0.1273	0.4704
5	1.9147	1.6635	–0.6453
6	–0.1059	0.5536	–0.7659
7	–0.6093	–0.3770	–0.7659
8	–0.2116	–0.1341	0.7117
9	–0.7196	2.9187	–0.8866
10	–0.3311	0.4742	–0.7358
11	2.7285	–0.7856	0.9529
12	–0.4897	0.3766	0.2593
13	–0.1059	–0.8650	0.7720
14	–0.4645	1.1642	–0.5850
15	0.8918	–0.8105	–0.7056
16	–0.3840	–0.9059	–0.7358
17	–0.0208	–0.8128	1.6465
18	–0.5817	–0.3838	–0.7659
19	–0.4139	0.4515	0.5910
20	–0.6438	–0.2295	–0.7659

At last, the $\boldsymbol{m}''_{ij}$ must be in interval [0, 1], the impact of dimension is avoided.

$$M'' = \begin{pmatrix} m''_{11} & m''_{12} & m''_{13} & \cdots & m''_{1j} \\ m''_{21} & m''_{22} & m''_{23} & & \vdots \\ m''_{31} & m''_{32} & m''_{33} & & \vdots \\ \vdots & & & \ddots & \vdots \\ m''_{i1} & \cdots & \cdots & \cdots & m''_{ij} \end{pmatrix}$$

4.2. **Creating fuzzy similar matrix.** The form of similar relation matrix (degree of membership matrix) $\boldsymbol{R}$ is shown:

$$R = \begin{pmatrix} r_{11} & & & & \\ r_{21} & r_{22} & & & \\ r_{31} & r_{32} & r_{33} & & \\ \vdots & & & \ddots & \\ r_{n1} & \cdots & \cdots & \cdots & r_{nn} \end{pmatrix}$$

The relation of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is the same as the relation of $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$ ($i, j = 1, 2, 3, \ldots, n$), so we just need half of matrix which is divided by diagonal. In addition, any element $\boldsymbol{x}_i$

TABLE 4. Standard data transformation

No.	Attribute		
	Economic loss	**Disaster area**	**Casualty state**
	X_1(mil.\$)	X_2(m^2)	X_3
1	0.0393	0.0504	0.3000
2	0.7143	0.0136	1.0000
3	0.0000	0.0789	0.0000
4	0.0644	0.2036	0.3833
5	0.7722	0.6718	0.0750
6	0.2066	0.3816	0.0417
7	0.0656	0.1383	0.0417
8	0.1770	0.2018	0.4500
9	0.0347	1.0000	0.0083
10	0.1435	0.3608	0.0500
11	1.0000	0.0315	0.5167
12	0.0991	0.3353	0.3250
13	0.2066	0.0107	0.4667
14	0.1062	0.5412	0.0917
15	0.4858	0.0249	0.0583
16	0.1287	0.0000	0.0500
17	0.2304	0.0243	0.7083
18	0.0734	0.1365	0.0417
19	0.1203	0.3549	0.4167
20	0.0560	0.1769	0.0417

is the same as itself. So the form of similar relation matrix $\boldsymbol{R}$ is shown:

$$R = \begin{pmatrix} 1 & & & & \\ r_{21} & 1 & & & \\ r_{31} & r_{32} & 1 & & \\ \vdots & & & \ddots & \\ r_{n1} & \cdots & \cdots & r_{n(n-1)} & 1 \end{pmatrix}$$

where if the $\boldsymbol{r}_{ij}$ is "1", $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are exactly the same, or else if the $\boldsymbol{r}_{ij}$ is the "0", $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are exactly the different. In here, we use max-min method to calculate the $\boldsymbol{r}_{ij}$ and it is shown:

$$r_{ij} = \frac{\sum_{h=1}^{k} \min(m''_{ih}, m''_{jh})}{\sum_{h=1}^{k} \max(m''_{ih}, m''_{jh})} \tag{3}$$

where $i < j$, because we just need half of matrix which is divided by diagonal.

4.3. **Fuzzy clustering algorithm.** In the graphic algorithm which is structured by Clustering Analysis of Maximal Tree method, all of the objects are vertexes. If $\boldsymbol{r}_{ij} \neq 0$, vertex $\boldsymbol{i}$ and vertex $\boldsymbol{j}$ can be connected by a line until the all vertexes be connected. However, they cannot produce any circuit, because any two vertexes have one relation. So we need remove lines of minimum value which are in produced circuit. At last, we get a Maximal Tree which each side has a weight "$\boldsymbol{r}_{ij}$". By the Threshold-λ, we remove all the side which weight $\boldsymbol{r}_{ij} < \lambda$. In the remaining vertexes, any connected vertexes are the same cluster. So if we want to extract something, we need embed standard value of them

in data set. If any vertex can be clustered to a group with the test data according to the threshold λ, it must be a useful data element which we want. The detail of this analysis is explained in Section 5.

5. **The Application of Data Fuzzy Clustering.**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1																			
2	0.20	1																		
3	0.12	0.01	1																	
4	0.60	0.24	0.12	1																
5	0.09	0.33	0.05	0.19	1															
6	0.15	0.12	0.13	0.32	0.41	1														
7	0.26	0.07	0.32	0.37	0.16	0.39	1													
8	0.47	0.33	0.10	0.78	0.24	0.40	0.30	1												
9	0.07	0.02	0.08	0.17	0.39	0.34	0.16	0.15	1											
10	0.17	0.10	0.14	0.36	0.36	0.86	0.44	0.40	0.34	1										
11	0.24	0.61	0.02	0.28	0.40	0.15	0.08	0.38	0.03	0.12	1									
12	0.51	0.21	0.10	0.73	0.29	0.52	0.32	0.65	0.27	0.58	0.25	1								
13	0.48	0.40	0.01	0.52	0.15	0.25	0.15	0.73	0.03	0.20	0.44	0.43	1							
14	0.19	0.09	0.11	0.35	0.47	0.63	0.33	0.34	0.49	0.67	0.11	0.54	0.17	1						
15	0.15	0.32	0.04	0.14	0.37	0.30	0.19	0.23	0.04	0.24	0.37	0.16	0.28	0.17	1					
16	0.19	0.10	0.00	0.16	0.12	0.27	0.34	0.22	0.04	0.32	0.12	0.19	0.26	0.21	0.31	1				
17	0.37	0.55	0.02	0.41	0.15	0.21	0.12	0.57	0.03	0.17	0.44	0.35	0.71	0.15	0.26	0.19	1			
18	0.26	0.07	0.31	0.37	0.17	0.40	0.96	0.30	0.16	0.45	0.09	0.33	0.16	0.34	0.21	0.36	0.13	1		
19	0.44	0.27	0.09	0.73	0.30	0.51	0.28	0.75	0.26	0.57	0.30	0.85	0.53	0.51	0.16	0.19	0.43	0.28	1	
20	0.25	0.06	0.29	0.42	0.18	0.44	0.83	0.33	0.20	0.50	0.08	0.36	0.13	0.37	0.17	0.27	0.11	0.80	0.31	1

FIGURE 1. Fuzzy similar data sheet

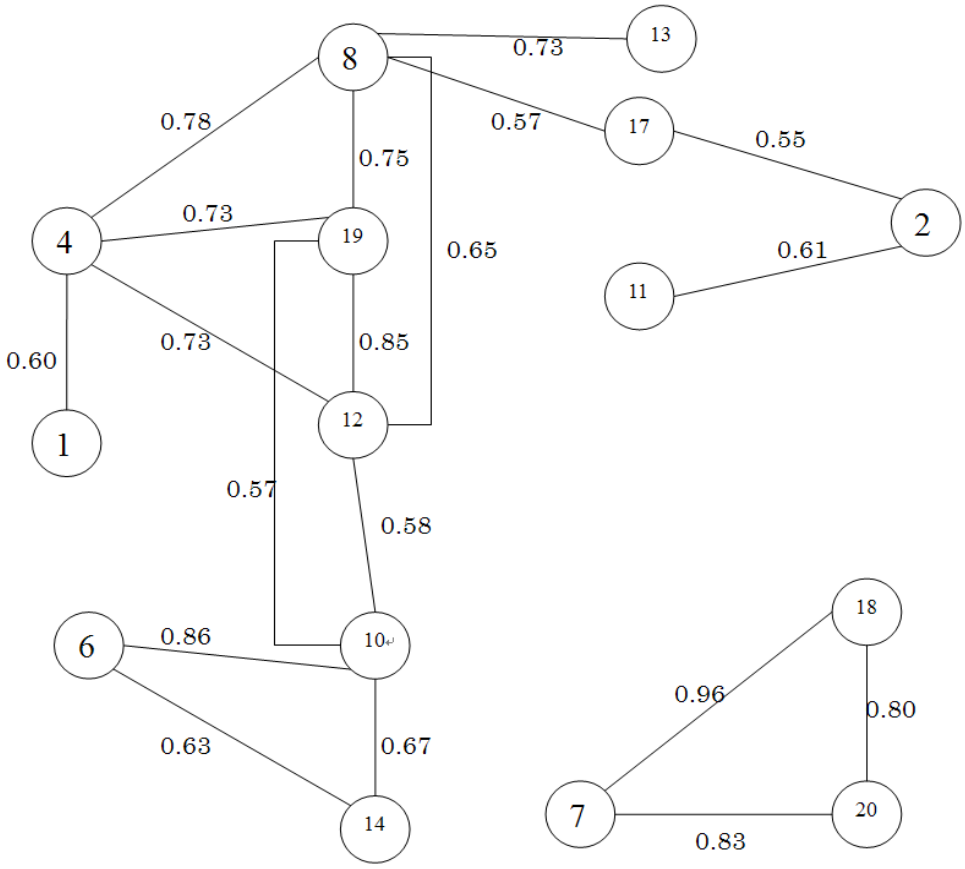

FIGURE 2. Cluster tree in the primary

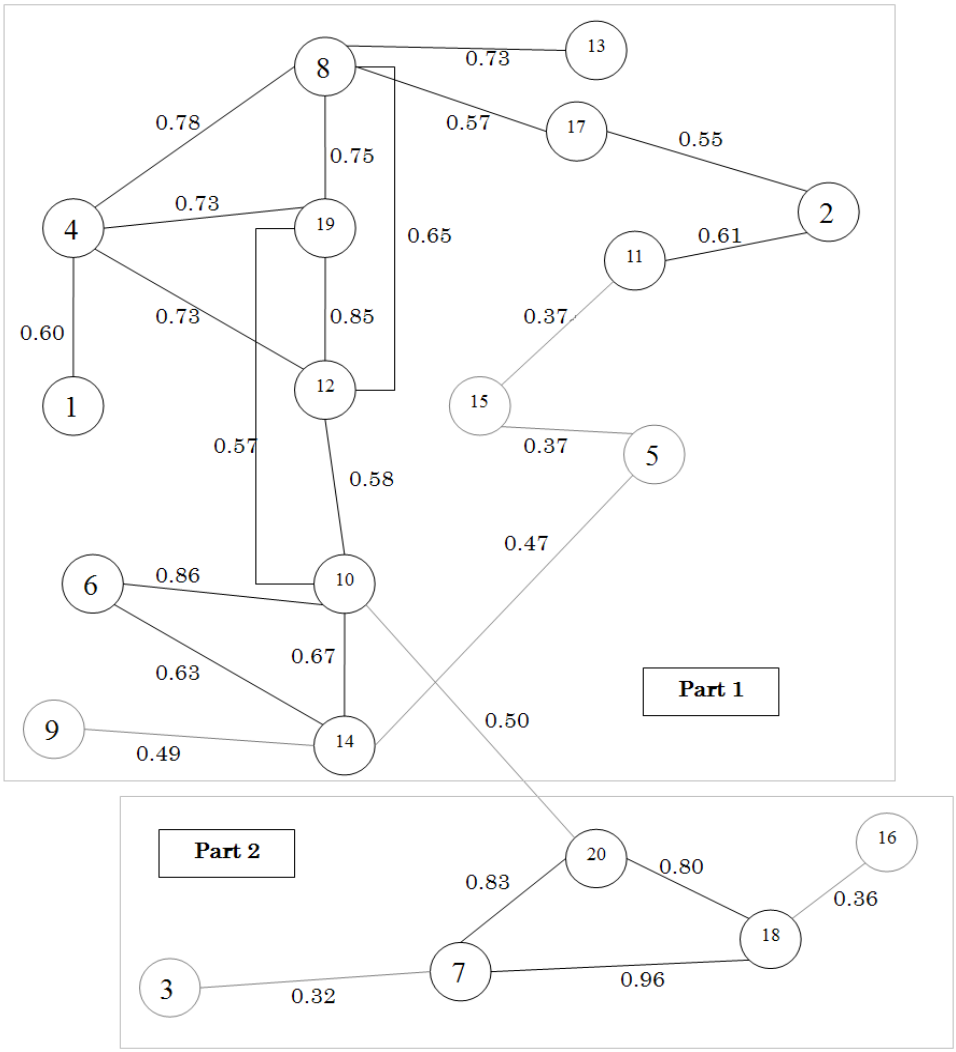

FIGURE 3. Secondary cluster tree

5.1. **The production process of maximal tree.** Let us illustrate our model by applying it to a data mining system of accident damage assessment. There are three types of attribute which are Economic loss, Disaster area, Casualty state for evaluating severity of accident. There are 18 example data is in Table 1.

This time we want to get two type cases which the one like $ 200 mil. is for Economic loss state, 600m^2 is for Disaster area state and 50 people is for Casualties state, and the other one like $ 100 mil. is for Economic loss state, 300m^2 is for Disaster area state and 5 people for Casualties state. The new data set in Table 2.

In Table 2, No.19 and No.20 are the evaluation data. We use fuzzy cluster method to extract all data. The detail is in the following:

Step 1:

Using Equation (1) to transform data which is in Table 2 and obtain standard deviation transformation data in Table 3.

Step 2:

Using Equation (2) to transform data which is in Table 3 and obtain range transformation data in Table 4.

Step 3:

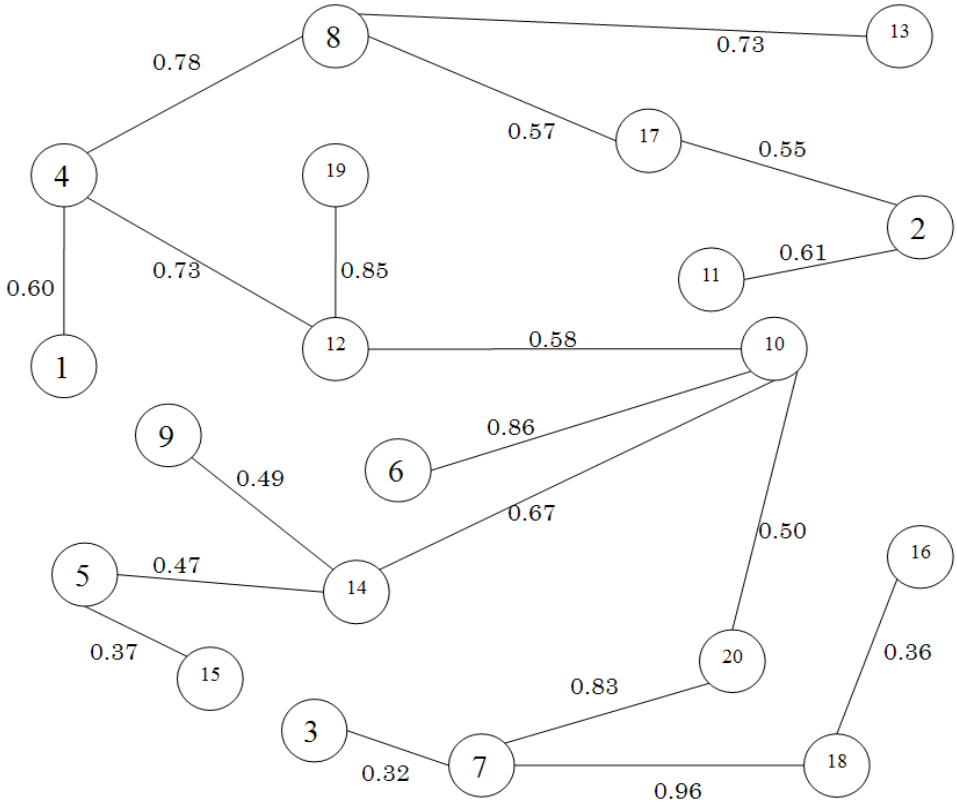

FIGURE 4. The maximal tree

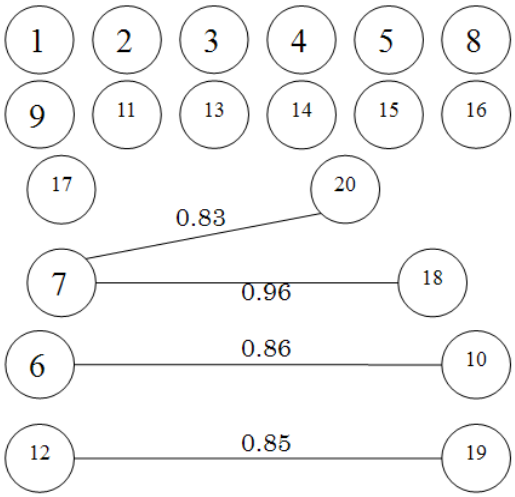

FIGURE 5. The clustering tree ($\lambda = 0.8$)

We use max-min method to calculate the $\boldsymbol{r}_{ij}$ and it is shown in Figure 1. Any element $\boldsymbol{x}_i$ is the same as itself, so the all diagonal element are "1".

Step 4:

At last, we use fuzzy similar data to create maximal tree method. The all are 20 vertexes. So we pick up 20 pair vertexes ({1, 4}, {2, 11}, {2, 17}, {4, 8}, {4, 12}, {4, 19}, {6, 10}, {6, 14}, {7, 18}, {7, 20}, {8, 12}, {8, 13}, {8, 17}, {8, 19}, {10, 12}, {10, 14},

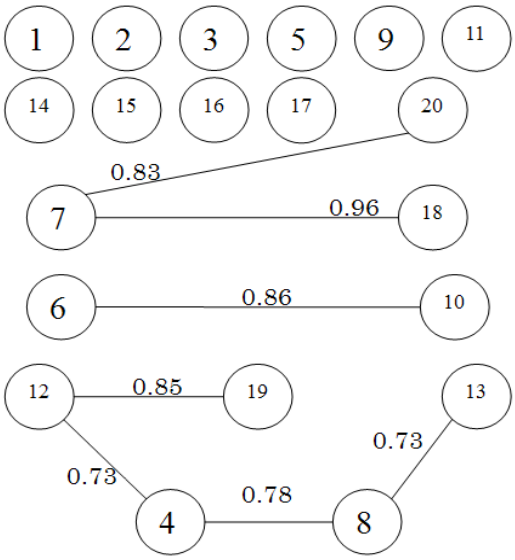

FIGURE 6. The clustering tree ($\lambda = 0.7$)

{10, 19}, {12, 19}, {13, 17}, {18, 20}), which have maximum similar values and create cluster tree in the primary. It is shown as Figure 2.

Step 5:

But vertex 3, 5, 9, 15, 16 are not included in cluster tree. And then we need add these 5 vertexes and connect vertexes with maximum similar value of them. The similar values 0.32 is for {3, 7}, 0.47 is for {5, 14}, 0.49 is for {9, 14}, 0.37 is for {15, 5}&{15, 11}, and 0.36 is for {16, 18}. Of course, we need part 1 and part 2 with maximum similar values 0.50 by {10, 20}. Figure 3 is secondary cluster tree.

Step 6:

The Maximal Tree cannot produce any circuit, because any two vertexes have one relation. So we need remove sides of minimum value which are in produced circuit. The line {4, 19}, {8, 19}, {8, 12}, {10, 19}, {6, 14}, {18, 20}, {11, 15} would be removed. The last Maximal Tree is shown as Figure 4.

5.2. **The data mining of fuzzy clustering.** By the evaluation values of similar degree λ, if we want to extract accident case of high similar degree, we can set $\lambda = 0.8$. After remove all the side which its similar value below 0.8, we can get clustering like Figure 5. One group consisted case 7, 18 and 20, one group consisted case 6 and 10, one group consisted case 12 and 19, the other cases were 1, 2, 3, 4, 5, 8, 9, 11, 13, 14, 15, 16 and 17 respectively. The case 12 has the same group with evaluation case 19. And the case 7, 18 has the same group with evaluation case 20. So case 12 which is similar to case 19, and case 7, 18 are similar to case 20 are required for us.

If $\lambda = 0.7$, by the clustering tree (Figure 6), case 4, 8, 12, 13 which are similar to case 19, and case 7, 18 are similar to case 20 are required for us.

6. **Conclusions.** Fuzzy clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. So we can use this clustering method to identify useful information for us.

Fuzzy clustering is the process of dividing data elements into clusters so that items in the same clusters are as similar as possible while items across different clusters are as

dissimilar as possible. We can use this clustering method to cluster the data elements which we want to do.

A method to supervise the process of fuzzy clustering for information extraction in order to detect and remove less important clusters has been presented. The reduction is based on the Maximal Tree approach to subset selection and adopted for fuzzy clustering in this paper. The method is applicable for obtaining fuzzy rules from data for function approximation and systems modeling purposes. It helps the user in the difficult task of data mining and data classification when applying fuzzy clustering. This threshold is used by the algorithm for selecting the appropriate cluster for the considered data. The considered synthetic and real-world examples demonstrated the improved mining properties due to the befitting cluster and the algorithms capability of determining a suitable similar degree of clusters in the data.

REFERENCES

[1] R. A. Cuninghame-Green, Minimax algebra, *Lecture Notes in Economics and Mathematical Systems*, vol.166, 1979.

[2] M. Allame and B. Vatankhahan, Iteration algorithm for solving Ax = b in max-min algebra, *Appl. Math. Compute.*, vol.175, pp.269-276, 2006.

[3] K. Cechlárová, Unique solvability of max-min fuzzy equations and strong regularity of matrices over fuzzy algebra, *Fuzzy Sets and Systems*, vol.75, pp.165-177, 1995.

[4] Y. Matsumoto and J. Watada, Knowledge acquisition from time series data through rough sets analysis, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12(B), pp.4885-4897, 2009.

[5] A. Ghodousian and E. Khorram, Solving a linear programming problem with the convex combination of the max-min and the max-average fuzzy relation equations, *Appl. Math. Comput.*, vol.180, pp.411-418, 2006.

[6] *http://en.wikipedia.org/wiki/Fuzzy_set.*

[7] *http://en.wikipedia.org/wiki/Fuzzy_clustering.*