Computers and Mathematics with Applications 60 (2010) 151–165

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A robust clustering procedure for fuzzy data

Wen-Liang Hung [a], Miin-Shen Yang [b,*], E. Stanley Lee [c]

[a] Graduate Institute of Computer Science, National Hsinchu University of Education, Hsin-Chu, Taiwan

[b] Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan

[c] Department of Industrial and Manufacturing Systems Engineering, Kansas State University, KS 66506, USA

ARTICLE INFO

Article history:
Received 12 February 2010
Accepted 23 April 2010

Keywords:
Robust clustering algorithm
LR-type fuzzy number
Outlier
Robustness
Similarity measure
Tea evaluation

ABSTRACT

In this paper we propose a robust clustering method for handling *LR*-type fuzzy numbers. The proposed method based on similarity measures is not necessary to specify the cluster number and initials. Several numerical examples demonstrate the effectiveness of the proposed robust clustering method, especially robust to outliers, different cluster shapes and initial guess. We then apply this algorithm to three real data sets. These are Taiwanese tea, student data and patient blood pressure data sets. Because tea evaluation comes under an expert subjective judgment for Taiwanese tea, the quality levels are ambiguity and imprecision inherent to human perception. Thus, *LR*-type fuzzy numbers are used to describe these quality levels. The proposed robust clustering method successfully establishes a performance evaluation system to help consumers better understand and choose Taiwanese tea. Similarly, *LR*-type fuzzy numbers are also used to describe data types for student and patient blood pressure data. The proposed method actually presents good clustering results for these real data sets.

1. Introduction

Cluster analysis is an important tool for data analysis. It is a method for finding groups within data with most similar groups in the same cluster and most dissimilar groups between different clusters. The hierarchical clustering was supposed as an earliest clustering method used by biologist and social scientists. Afterwards, cluster analysis becomes a branch in statistical multivariate analysis so that many theories and various methods were investigated. On the other hand, the learning and recognition is generally beginning from clustering. In this case, cluster analysis becomes an unsupervised learning in pattern recognition. Since Zadeh [1] proposed fuzzy sets that produced the idea of partial memberships described by membership functions with allowing membership degrees to all clusters, it is very successfully used in cluster analysis. Up to date, fuzzy clustering has been widely studied and applied in a variety of substantive areas (see [2–5]). In fuzzy clustering, the fuzzy c-means (FCM) algorithm and its variations are the most well-known and used methods.

Fuzzy data is a data type with imprecision or with a source of uncertainty caused not by randomness, but by fuzziness, such as in linguistic assessments. Real fuzzy data types can be found in natural language, the social sciences and in knowledge representation. In general, *LR*-type fuzzy numbers (FNs) are convenient and useful to describe fuzzy data [6]. Hathaway et al. [7] proposed FCM clustering for symmetric trapezoidal FNs using a parametric approach. Yang and Ko [8] proposed a fuzzy c-numbers (FCN) clustering algorithm for handling *LR*-type FNs. But, these methods are always influenced by noise and outliers and often get improper clustering results when the data set includes different cluster sample sizes. To overcome these drawbacks, Hung and Yang [9] proposed the so-called alternative fuzzy c-numbers (AFCN) clustering algorithm for

[*] Corresponding author. Tel.: +886 3 2653119.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

0898-1221/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2010.04.042

LR-type FNs by using an exponential-type distance measure. According to their results, AFCN is superior in performance to FCN. Recently, D'Urso and Giordani [10] defined a weighted distance between symmetric fuzzy data and used a weighted distance to modify the FCM clustering algorithm for fuzzy data. However, there are two main drawbacks in these algorithms. One is the clustering performance is sensitive to initial guess. Another one is the cluster number needs to be specified a priori.

To solve these drawbacks, we propose a clustering procedure that is robust to initials and cluster number. We will employ the method called "similarity-based clustering method (SCM)" proposed by Yang and Wu [11] to obtain better clustering results. The SCM clustering can self-organize the cluster number and structure for numerical data. In this paper, we modify the method such that it could handle *LR*-type FNs. In the parameter estimation step, we use a graphical method of correlation comparisons to replace the original correlation comparison algorithm of Yang and Wu [11] for finding a suitable parameter γ. As a whole, the proposed method can be robust to initials, cluster number, cluster shapes and outliers for handling *LR*-type fuzzy data.

The remainder of this paper is organized as follows. In Section 2, the modification of SCM for fuzzy data is first described and then a suitable parameter γ is selected. We use a graphical method of correlation comparisons to select a suitable parameter γ. We then create a robust clustering algorithm for *LR*-type FNs. In Section 3, several numerical examples are presented to illustrate the effectiveness of the proposed algorithm. Since tea has long been an important agricultural product in Taiwan and its type and price are wide ranging and complex, to establish an evaluation system for consumers better understanding Taiwanese tea is an important issue. However, the problem of establishing an evaluation system can be considered as a clustering process to group the related various criteria together. In addition, due to the vagueness of human feelings and recognition, fuzzy data are better used to describe the quality levels of criteria. In Section 4, we will apply the proposed algorithm to create a performance evaluation of Taiwan tea. Moreover, we also apply the proposed algorithm for clustering student and patient blood pressure data. Finally, conclusions will be stated in Section 5.

2. A robust clustering algorithm for fuzzy data

We propose a robust clustering method for fuzzy data in this section. To consider a clustering method for handling *LR*-type FNs, we need to have a distance measure between the two *LR*-type FNs. Let L (and R) be decreasing, shape functions from $\Re^+ = [0, \infty)$ to $[0, 1]$ with $L(0) = 1$; $L(x) < 1$ for all $x > 0$; $L(x) > 0$ for all $x < 1$; $L(1) = 0$ or ($L(x) > 0$ for all x and $L(+\infty) = 0$). An FN X is called *LR*-type (see [12] pp. 62–63) if for $m \in \Re = (-\infty, \infty)$, $\alpha > 0$, $\beta > 0$,

$$X(x) = \begin{cases} L\left(\dfrac{m-x}{\alpha}\right) & \text{for } x \le m, \\ R\left(\dfrac{x-m}{\beta}\right) & \text{for } x \ge m, \end{cases}$$

where m is called the mean value of X and α and β are called the left and right spreads respectively. Symbolically X is denoted by $(m, \alpha, \beta)_{LR}$. Then, a distance for any $X = (m_X, \alpha_X, \beta_X)_{LR}$ and $Y = (m_Y, \alpha_Y, \beta_Y)_{LR}$ is defined as follows (see [8]):

$$d^2(X, Y) = \frac{1}{3}\{(m_X - m_Y)^2 + [(m_X - l\alpha_X) - (m_Y - l\alpha_Y)]^2 + [(m_X + r\beta_X) - (m_Y + r\beta_Y)]^2\}$$

where $l = \int_0^1 L^{-1}(w)\mathrm{d}w$ and $r = \int_0^1 R^{-1}(w)\mathrm{d}w$. In *LR*-type FNs, the triangular fuzzy numbers (TFNs) are most commonly used. For a *LR*-type FN $X = (m, \alpha, \beta)_{LR}$, if $L(x) = R(x) = 1 - x$ then X is called a TFN, denoted by $X = (m, \alpha, \beta)_T$, i.e.

$$X(x) = \begin{cases} 1 - \dfrac{m-x}{\alpha} & \text{for } x \le m (\alpha > 0), \\ 1 - \dfrac{x-m}{\beta} & \text{for } x \ge m (\beta > 0). \end{cases}$$

Obviously, $r = l = \frac{1}{2}$ in $d^2(X, Y)$ for two TFNs $X = (m_X, \alpha_X, \beta_X)_T$ and $Y = (m_Y, \alpha_Y, \beta_Y)_T$.

Let $\mathcal{G} = \{X_1, \ldots, X_n\}$ be a set of *LR*-type FNs with $X_j = (m_{X_j}, \alpha_{X_j}, \beta_{X_j})_{LR}$, $j = 1, \ldots, n$ and let c be the specified cluster number. Our goal is to cluster $\mathcal{G}$ into c clusters. Let $Z_i = (m_{Z_i}, \alpha_{Z_i}, \beta_{Z_i})_{LR}$ be the ith cluster center. Adopting Zadeh's idea [13], the following similarity measure $S(X_j, Z_i)$ between X_j and Z_i is used:

$$S(X_j, Z_i) = \exp\left(-\frac{d^2(X_j, Z_i)}{\sigma^2}\right), \tag{1}$$

where the σ^2 is the normalized constant defined by

$$\sigma^2 = \frac{\sum_{j=1}^{n} d^2(X_j, \bar{X})}{n} \quad \text{with } \bar{X} = (m_{\bar{X}}, \alpha_{\bar{X}}, \beta_{\bar{X}})_{LR}, \tag{2}$$

$$m_{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} m_{X_j}, \qquad \alpha_{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \alpha_{X_j}, \qquad \beta_{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \beta_{X_j}.$$

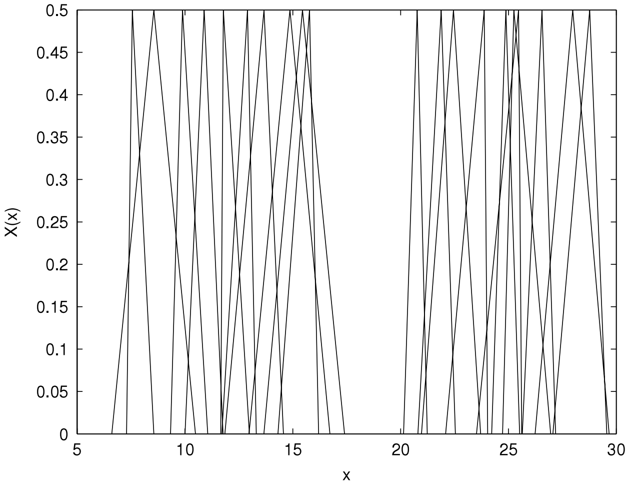

Fig. 1. The data set with two well-separated clusters.

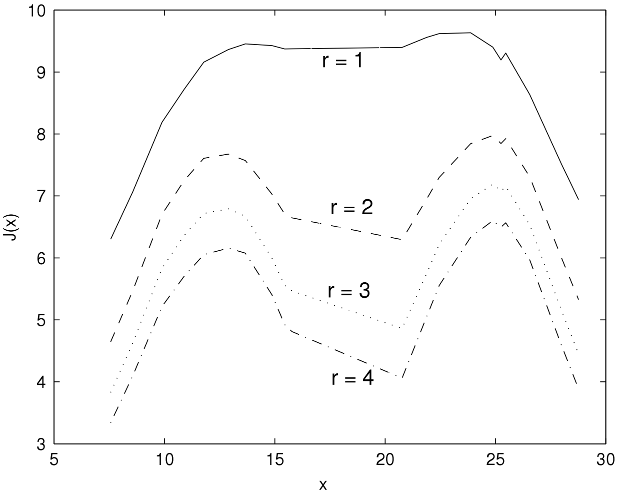

Fig. 2. The density shapes with different γ.

Then the total similarity measure is given by (cf. [11])

$$J(Z) = \sum_{i=1}^{c} \sum_{j=1}^{n} \left(\exp\left(-\frac{d^2(X_j, Z_i)}{\sigma^2} \right) \right)^{\gamma}, \tag{3}$$

where $\gamma > 0$. The role of γ is similar to the bandwidth in the kernel estimator. Increasing γ is equivalent to decrease the bandwidth. Thus, larger γ yields more peaks in the objective function $J(Z)$, i.e., more clusters. Later we will discuss the effect of γ on the clustering results.

In this paper, we use a graphical method of correlation comparisons to find the suitable γ which is different from the CCA of [11]. Considering the following function

$$\tilde{J}(X_k)_{\gamma} = \sum_{j=1}^{n} \left(\exp\left(-\frac{d^2(X_j, X_k)}{\sigma^2} \right) \right)^{\gamma} \quad k = 1, \ldots, n, \tag{4}$$

this function represents the density shape of the data points and is similar to the mountain function proposed by Yager and Filev [14]. We use the data set with triangle fuzzy numbers, shown in Fig. 1 to explain how to use Eq. (4) to find a suitable γ. Fig. 1 clearly indicates that there are two well-separated clusters. Thus, the density shape of the data points should have two peaks. From Fig. 2, the curves with $\gamma = 2, 3, 4$ also have two peaks. Furthermore, the shapes are unchanged with $\gamma = 3$ and $\gamma = 4$, the correlation coefficient between $\{\tilde{J}(X_k)_{\gamma=3} | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma=4} | k = 1, \ldots, n\}$ is very close to 1. Therefore, $\gamma = 3$ is a suitable parameter estimate. The following example is illustrated the graphical method of correlation comparisons to find the suitable γ.

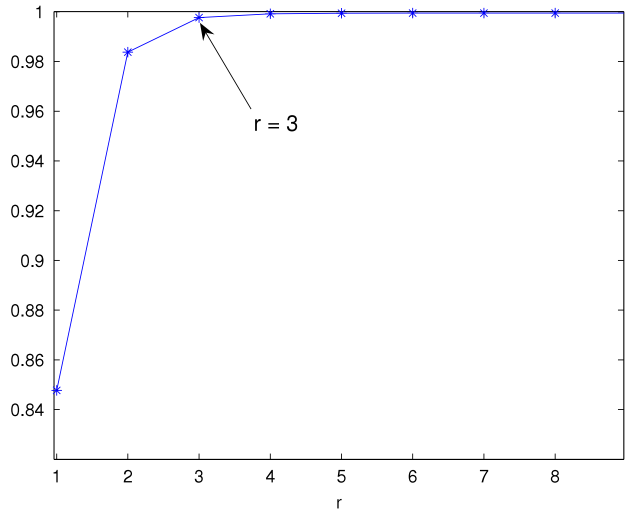

Fig. 3. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \ldots, n\}$ for Example 1.

Example 1. We consider the data set given in Fig. 1. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \ldots, n\}$ are shown in Fig. 3. In Fig. 3, the y-coordinate of the first star point represents the correlation coefficient between $\gamma = 1$ and $\gamma = 2$. The second star point represents the correlation coefficient between $\gamma = 2$ and $\gamma = 3$. The third star point represents the correlation coefficient between $\gamma = 3$ and $\gamma = 4$, etc. Since the density shape will be unchanged when the correlation coefficient is close to 1, we may choose γ with the star point close to the line of value 1. Note that the shapes are unchanged with $\gamma = 3$ and $\gamma = 4$ for the aforementioned Fig. 2. It is seen that $\gamma = 3$ is a suitable estimate as shown in Fig. 3.

After the parameter γ is estimated by the previous approach, the next step is to find Z_i that maximize the objective function $J(Z)$. Differentiating $J(Z)$ with respect to all $Z_i = (m_{Z_i}, \alpha_{Z_i}, \beta_{Z_i})_{LR}$, we obtain the following necessary conditions:

$$m_{Z_i} = \frac{\sum_{j=1}^{n} S_{ij}^{\gamma}[3m_{X_j} + l(\alpha_{Z_i} - \alpha_{X_j}) + r(\beta_{X_j} - \beta_{Z_i})]}{3\sum_{j=1}^{n} S_{ij}^{\gamma}}, \quad i = 1, \ldots, c, \tag{5}$$

$$\alpha_{Z_i} = \frac{\sum_{j=1}^{n} S_{ij}^{\gamma}(m_{Z_i} - m_{X_j} + l\alpha_{X_j})}{l\sum_{j=1}^{n} S_{ij}^{\gamma}}, \quad i = 1, \ldots, c, \tag{6}$$

$$\beta_{Z_i} = \frac{\sum_{j=1}^{n} S_{ij}^{\gamma}(m_{X_j} - m_{Z_i} + r\beta_{X_j})}{r\sum_{j=1}^{n} S_{ij}^{\gamma}}, \quad i = 1, \ldots, c, \tag{7}$$

where

$$S_{ij} = \exp\left(-\frac{d^2(X_j, Z_i)}{\sigma^2}\right). \tag{8}$$

Note that m_{Z_i}, α_{Z_i} and β_{Z_i} in Eqs. (5)–(7) cannot be solve directly. However, the fixed-point iterative method can be used to numerically solve them. This forms the similarity clustering algorithm (SCA). After the graphical method of correlation comparisons gets an estimate γ, the SCA will be used to find the peaks of the objective function (3). The SCA algorithm for *LR*-type FNs can be summarized as follows:

SCA algorithm for LR-type FNs

(S1) Fix any $\epsilon > 0$. Choose initial mean value, left and right spreads, say $Z_i^{(0)} = (m_{Z_i}^{(0)}, \alpha_{Z_i}^{(0)}, \beta_{Z_i}^{(0)}), i = 1, \ldots, c$.

(S2) Calculate $S_{ij}^{(0)}$ using $Z_i^{(0)}$ and Eq. (8).

(S3) Update $Z_i^{(0)}$ by $Z_i^{(1)}$ using $S_{ij}^{(0)}$ and Eqs. (5)–(7).

(S4) IF $\max_i d^2(Z_i^{(0)}, Z_i^{(1)}) < \epsilon$, stop. Otherwise, set $Z_i^{(1)} = Z_i^{(0)}$ and GO TO (S2).

We mention that the value of γ determines the number of peaks in the objective function $J(Z)$. The reasons are below. Let

$$\lim_{\gamma\to 0} m_{Z_i} = m^*_{Z_i}, \qquad \lim_{\gamma\to 0} \alpha_{Z_i} = \alpha^*_{Z_i}, \qquad \lim_{\gamma\to 0} \beta_{Z_i} = \beta^*_{Z_i}.$$

Then

$$\lim_{\gamma\to 0} m_{Z_i} = \frac{\sum_{j=1}^{n}[3m_{X_j} + l(\alpha^*_{Z_i} - \alpha_{X_j}) + r(\beta_{X_j} - \beta^*_{Z_i})]}{3n}, \tag{9}$$

$$\lim_{\gamma\to 0} \alpha_{Z_i} = \frac{1}{l} m^*_{Z_i} - \frac{1}{l} m_{\bar{X}} + \alpha_{\bar{X}}, \tag{10}$$

$$\lim_{\gamma\to 0} \beta_{Z_i} = \frac{1}{r} m_{\bar{X}} - \frac{1}{r} m^*_{Z_i} + \beta_{\bar{X}}. \tag{11}$$

Substituting Eqs. (10) and (11) into Eq. (9), we obtain

$$\begin{aligned} m^*_{Z_i} &= m_{\bar{X}} + \frac{l}{3}(\alpha^*_{Z_i} - \alpha_{\bar{X}}) + \frac{r}{3}(\beta_{\bar{X}} - \beta^*_{Z_i}) \\ &= m_{\bar{X}} + \frac{1}{3}(m^*_{Z_i} - m_{\bar{X}}) + \frac{1}{3}(m^*_{Z_i} - m_{\bar{X}}) \\ &= \frac{1}{3} m_{\bar{X}} + \frac{2}{3} m^*_{Z_i}. \end{aligned}$$

This implies $m^*_{Z_i} = m_{\bar{X}}$. It follows $\alpha^*_{Z_i} = \alpha_{\bar{X}}$, $\beta^*_{Z_i} = \beta_{\bar{X}}$.

This illustrates that when $\gamma \to 0$, the objective function $J(Z)$ has only one peak on the sample mean $\bar{X} = (m_{\bar{X}}, \alpha_{\bar{X}}, \beta_{\bar{X}})_{LR}$. We point out what will happen as $\gamma \to \infty$ as follows. Let $S^*_{ij} = S_{ij}/S$, where $S = \max\{S_{i1}, \dots, S_{in}\}$. After simple calculations, we have

$$\lim_{\gamma\to\infty} m_{Z_i} = \frac{\sum_{S^*_{ij}=1} m_{X_j}}{\sum_{S^*_{ij}=1} 1}, \qquad \lim_{\gamma\to\infty} \alpha_{Z_i} = \frac{\sum_{S^*_{ij}=1} \alpha_{X_j}}{\sum_{S^*_{ij}=1} 1}, \qquad \lim_{\gamma\to\infty} \beta_{Z_i} = \frac{\sum_{S^*_{ij}=1} \beta_{X_j}}{\sum_{S^*_{ij}=1} 1}.$$

It means that when $\gamma \to \infty$, almost of data points are the peaks.

When one processes the SCA algorithm, all the cluster centers, Z_i, will change positions for each iteration. If the data set has only one peak on the objective function $J(Z)$, all the centers will gradually centralize to that unique peak. In this case, we will claim there is only one cluster for this data set. When the data set has more than one peak on the objective function $J(Z)$, we can randomly give more initial cluster centers to process this algorithm and these centers will then centralize to the peaks of the objective function $J(Z)$. The problem here is what kind of the initialization can guarantee that all peaks (clusters) will be found simultaneously. To solve this problem, we adopt Yang and Wu's [11] suggestion to set all data points to be the initial centers (i.e., $Z^{(0)} = (Z_1^{(0)}, \dots, Z_n^{(0)}) = (X_1, \dots, X_n)$).

Because SCA is processed with $Z^{(0)} = (Z_1^{(0)}, \dots, Z_n^{(0)}) = (X_1, \dots, X_n)$, the final n cluster centers will show the final states of all data points. This provides a method to classify the data points using their final states. For example, $Z_1^{(0)}$ and $Z_n^{(0)}$ converge to the same peak. Then $Z_1^{(0)}$ and $Z_n^{(0)}$ should belong to the same cluster and so should X_1 and X_n. Therefore, when $Z^{(0)} = (Z_1^{(0)}, \dots, Z_n^{(0)}) = (X_1, \dots, X_n)$ is centralized to c^* clusters, the data set will also be classified into those c^* clusters simultaneously. By processing the final states of $Z^{(0)} = (Z_1^{(0)}, \dots, Z_n^{(0)}) = (X_1, \dots, X_n)$ into the single linkage hierarchical algorithm, the optimal cluster number c^* and the identified clusters will be found simultaneously. Thus, the proposed robust clustering algorithm for *LR*-type FNs can be summarized as follows.

A robust clustering algorithm for LR-type FNs

Step 1. Use the graphical method of correlation comparison to select γ.

Step 2. Process SCA for *LR*-type FNs with $Z^{(0)} = (Z_1^{(0)}, \dots, Z_n^{(0)}) = (X_1, \dots, X_n)$.

Step 3. Process the hierarchical algorithm with the final states of the data points.

Step 4. Find the optimal cluster number c^* according to the dendrogram.

Step 5. Identify these c^* clusters.

3. Numerical examples

In this section we present some examples to illustrate the effectiveness of the proposed robust clustering algorithm. The data set in Example 2 is from [9] where there are 20 *LR*-type FNs with two clusters. To see the effects of outliers on

Table 1
Data set $\mathcal{G}_1$.

No.	1	2	3	4	5
$X = (m, \alpha, \beta)_T$	(7.56, 0.27, 1.00)	(8.56, 1.95, 1.93)	(9.89, 0.56, 1.17)	(10.89, 0.89, 0.88)	(11.78, 0.12, 1.21)
No.	6	7	8	9	10
$X = (m, \alpha, \beta)_T$	(12.90, 1.19, 0.41)	(13.67, 1.82, 0.90)	(14.87, 1.90, 1.85)	(15.45, 1.79, 1.95)	(15.78, 1.47, 0.42)
No.	11	12	13	14	15
$X = (m, \alpha, \beta)_T$	(20.77, 0.63, 0.47)	(21.88, 1.08, 0.66)	(22.45, 1.48, 1.26)	(23.88, 1.79, 0.16)	(24.88, 0.66, 0.64)
No.	16	17	18	19	20
$X = (m, \alpha, \beta)_T$	(25.25, 0.52, 1.71)	(25.47, 1.95, 0.15)	(26.56, 0.92, 0.63)	(27.98, 1.74, 1.69)	(28.77, 1.71, 0.79)

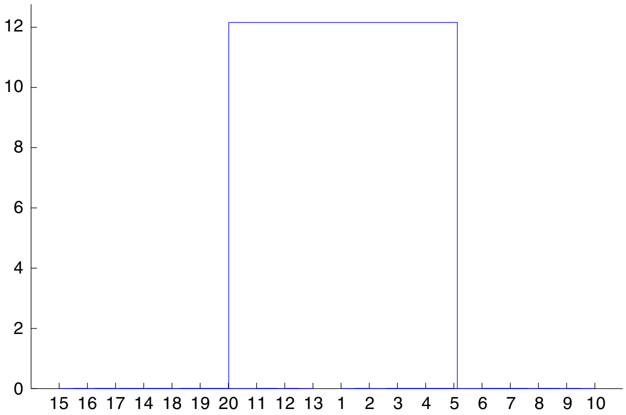

Fig. 4. The dendrogram through the proposed algorithm for Example 2.

the proposed robust clustering method for *LR*-type fuzzy numbers, we use Example 3 by adding two outlying points to the data set in Example 2. In Example 4, we consider a well-known clustering problem (cf. [15]) where there is an inordinate difference in the number of members in each sample cluster. We implement the robust clustering algorithm for these examples with $\epsilon = 0.0001$.

Example 2. In this example, we use the data set as shown in Fig. 1. We present these data points in Table 1. The data set are fuzzy data with two well-separated clusters. From Example 1, $\gamma = 3$ is a suitable estimate for this fuzzy data set. We then run the SCA algorithm for the data set and then use the single linkage hierarchical algorithm with the final positions of all data points to find the optimal cluster number c^*. The results are shown as the dendrogram in Fig. 4. This dendrogram clearly indicates that there are two well-separated clusters and hence the optimal cluster number $c^* = 2$. The corresponding cluster centers are $Z_1 = (12.797, 1.199, 1.081)_T$ and $Z_2 = (24.843, 1.230, 0.781)_T$. Therefore, the identified clusters are found that data numbers 1–10 belong to cluster 1 and data numbers 11–20 belong to cluster 2.

Example 3. To see the effect of outliers on the proposed clustering algorithm for fuzzy data, we add two points $(40, 1.82, 0.15)_T$ and $(50, 0.71, 1.79)_T$ to the data set $\mathcal{G}_1$. Fig. 5 shows that these added points are far away from the other fuzzy data so that they can be regarded as outliers. Based on the graphical method of correlation comparisons as shown in Fig. 6, $\gamma = 7$ is a suitable estimate. In the same way, we implement the SCA for the data set and then single linkage hierarchical algorithm on the final clustering data. The result is shown as the dendrogram in Fig. 7. This dendrogram clearly indicates that there are four well-separated clusters and hence the optimal cluster number $c^* = 4$. The corresponding cluster centers are: $Z_1 = (12.606, 1.188, 1.106)_T$, $Z_2 = (24.763, 1.236, 0.792)_T$, $Z_3 = (40.004, 1.818, 0.154)_T$ and $Z_4 = (49.983, 0.712, 1.787)_T$. Compared with the results of Example 2 without outliers, we find that the first two cluster centers keep the values of the cluster centers around 12 and 24. This reflects that the proposed robust clustering algorithm for fuzzy data is able to tolerate outliers. Furthermore, the proposed SCA for fuzzy data not only gives a perfect clustering result but also identifies outliers. We have the results that data numbers 1–10 belong to cluster 1 and numbers 11–20 belong to cluster 2 where these two added points are outliers.

Example 4 (*cf.* [9]). We consider a well-known clustering problem (cf. [15]) where there is an inordinate difference in the number of members in each sample cluster. We give the data set as shown in Table 2 where there is one large cluster (from data numbers 1 to 30) and one small cluster (from data numbers 31 to 35). The data set is also shown in Fig. 8. The corresponding ideal centers are $(22.947, 1.130, 1.082)_T$ and $(42.806, 1.368, 0.994)_T$, respectively. According to the graphical method of correlation comparisons shown in Fig. 9, $\gamma = 4$ is a suitable estimate. In the same way, we implement the SCA for the data set and then single linkage hierarchical algorithm on the final clustering data. The result is shown as the dendrogram in Fig. 10. This dendrogram clearly indicates that there are two well-separated clusters and

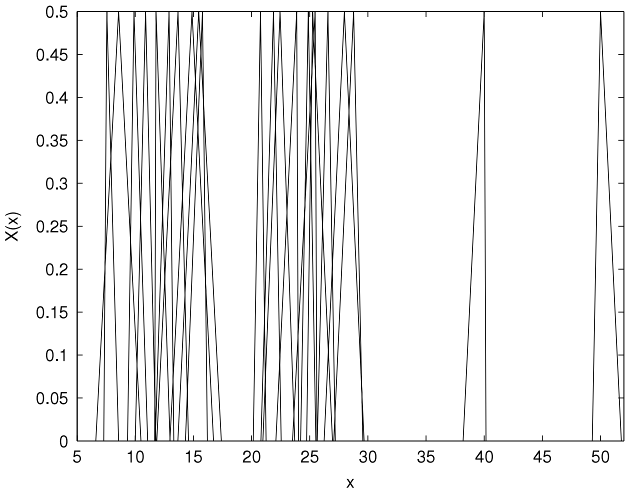

Fig. 5. The data set for Example 3.

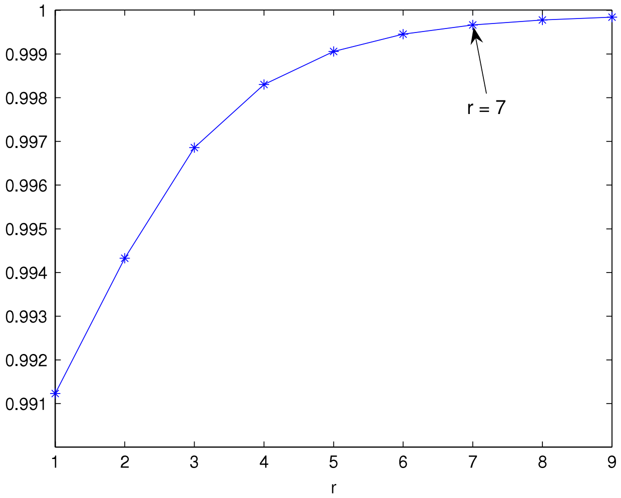

Fig. 6. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \ldots, n\}$ for Example 3.

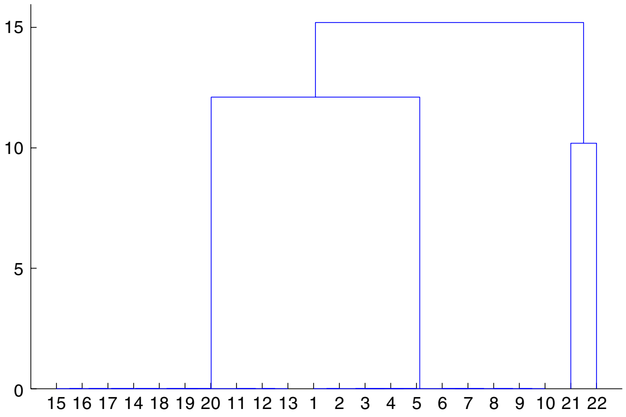

Fig. 7. The dendrogram through the proposed algorithm for Example 3.

hence the optimal cluster number $c^* = 2$. The corresponding cluster centers are: $Z_1 = (26.759, 1.171, 0.975)_T$ and $Z_2 = (42.741, 1.342, 1.018)_T$. Compared with the ideal centers, we find that the cluster centers are close to those ideal

Table 2
Data set for Fig. 8.

No.	1	2	3	4	5
$X = (m, \alpha, \beta)_T$	(9.56, 0.27, 1.00)	(10.56, 1.95, 1.93)	(10.89, 0.56, 1.17)	(13.89, 0.89, 0.88)	(14.78, 0.12, 1.21)
No.	6	7	8	9	10
$X = (m, \alpha, \beta)_T$	(14.90, 1.19, 0.41)	(15.67, 1.82, 0.90)	(16.87, 1.90, 1.85)	(17.45, 1.79, 1.95)	(19.78, 1.47, 0.42)
No.	11	12	13	14	15
$X = (m, \alpha, \beta)_T$	(21.02, 0.63, 1.92)	(22.56, 1.19, 0.41)	(22.89, 0.56, 1.17)	(23.78, 0.12, 1.21)	(23.92, 1.19, 0.41)
No.	16	17	18	19	20
$X = (m, \alpha, \beta)_T$	(24.67, 1.82, 0.90)	(25.13, 0.69, 0.64)	(25.92, 1.79, 1.95)	(26.02, 1.47, 0.42)	(26.30, 1.61, 1.92)
No.	21	22	23	24	25
$X = (m, \alpha, \beta)_T$	(27.77, 0.63, 0.47)	(28.08, 1.08, 0.66)	(28.90, 1.48, 1.26)	(29.00, 1.79, 0.16)	(29.67, 0.66, 0.64)
No.	26	27	28	29	30
$X = (m, \alpha, \beta)_T$	(30.77, 1.63, 1.47)	(31.08, 0.87, 1.66)	(31.90, 1.08, 0.68)	(32.00, 0.79, 1.16)	(32.67, 0.86, 1.63)
No.	31	32	33	34	35
$X = (m, \alpha, \beta)_T$	(40.25, 0.52, 1.71)	(40.47, 1.95, 0.15)	(43.56, 0.92, 0.63)	(43.98, 1.74, 1.69)	(45.77, 1.71, 0.79)

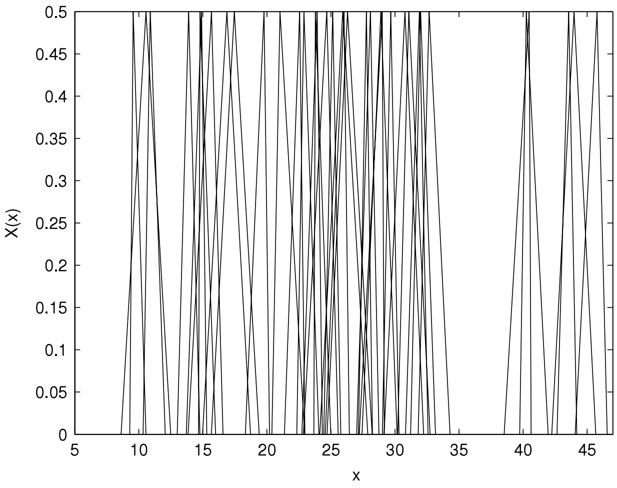

Fig. 8. The data set for Example 4.

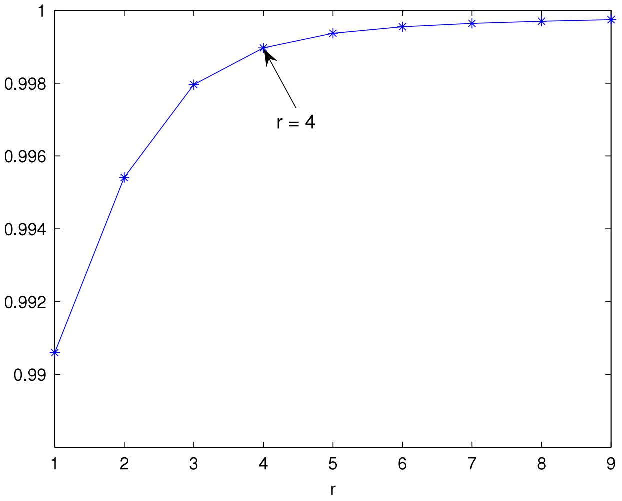

Fig. 9. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \dots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \dots, n\}$ for Example 4.

centers. Furthermore, the proposed algorithm for the data set well classifies these two clusters. That is, data numbers 1–30 belong to cluster 1 and data numbers 31–35 belong to cluster 2.

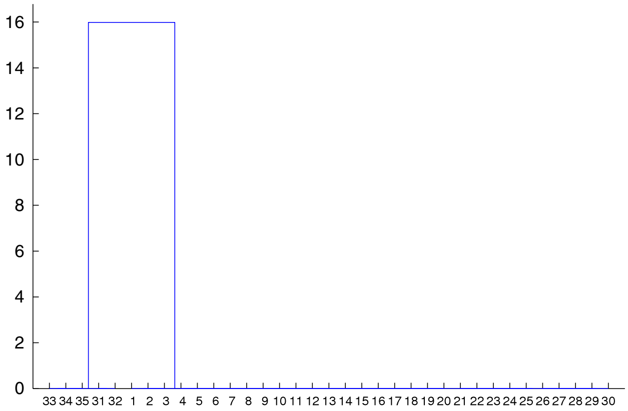

Fig. 10. The dendrogram through the proposed algorithm for Example 4.

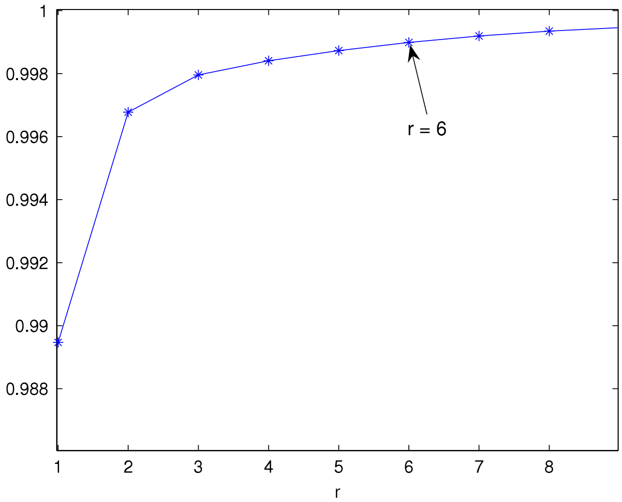

Fig. 11. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \ldots, n\}$ for data set $\mathcal{T}$.

4. Applications to real data sets

In this section, we apply the proposed robust clustering algorithm to real fuzzy data sets. We first consider a real example of Taiwanese tea from Hung and Yang [9].

Example 5. The Taiwanese tea data set has about 69 types of tea tree in Taiwan. The types of tea produced in Taiwan include green tea, Paochong, Oolong and black tea. In recent years, the majority of teas produced in Taiwan have been of the Paochong and Oolong varieties. Black tea and green tea are relatively minor types in comparison. Many consumers are confused due to the tea varieties and prices being numerous and complicated. To give consumers a better understanding of Taiwanese tea, the Taiwan Tea Experiment Station (TTES) is ongoing in its attempts to formulate an evaluation system for tea quality. This motivates us to develop an evaluation system by clustering method. In general, there are four criteria used to evaluate tea quality: appearance, tincture, liquid color and aroma and the quality levels are described using the terms: *perfect*, *good*, *medium*, *poor* and *bad*. The data are shown in Table A of Appendix.

In this data set, the overall performance $\bar{Y}_j$ for the jth type of tea is our interesting. A typical problem is "How to perform a clustering algorithm for $\bar{Y}_j$?". According to Hung and Yang [9], there are four criteria used to evaluate tea quality: appearance, tincture, liquid color and aroma and the quality levels are described using the terms: *perfect*, *good*, *medium*, *poor* and *bad*. These terms are described by TFNs as follows: $X_{perfect} = (1, 0.25, 0)_T$, $X_{good} = (0.75, 0.25, 0.25)_T$, $X_{medium} = (0.5, 0.25, 0.25)_T$, $X_{poor} = (0.25, 0.25, 0.25)_T$ and $X_{bad} = (0, 0, 0.25)_T$. The final evaluation data is shown in Table A. Let $Y_{jk} = (m_{Y_{jk}}, \alpha_{Y_{jk}}, \beta_{Y_{jk}})_T$ be assessed by the kth criterion for the jth type of tea, $k = 1, 2, 3, 4, j = 1, \ldots, n$.

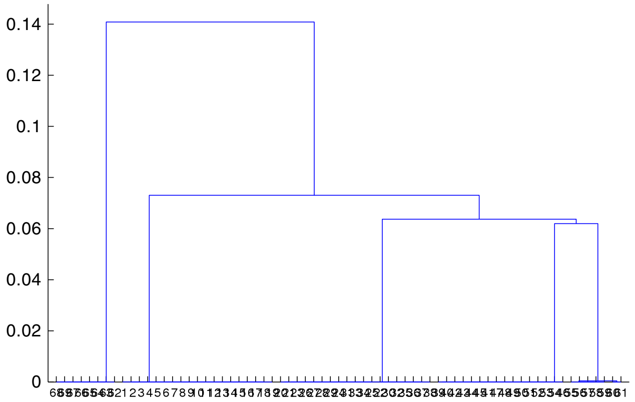

Fig. 12. The dendrogram through the proposed algorithm for data set $\mathcal{T}$.

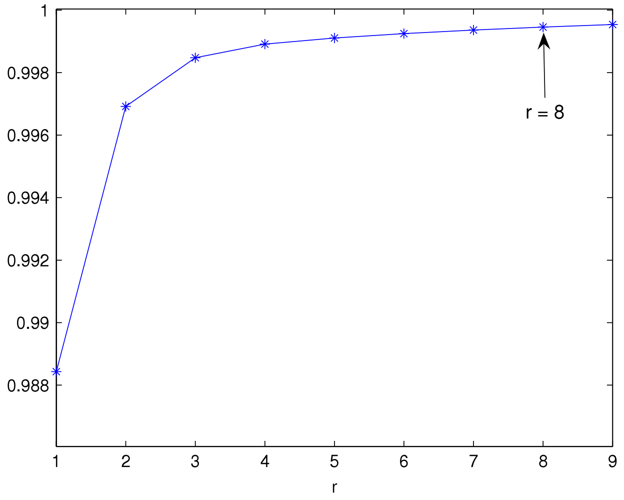

Fig. 13. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \ldots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \ldots, n\}$ for data set $\mathcal{T}_1$.

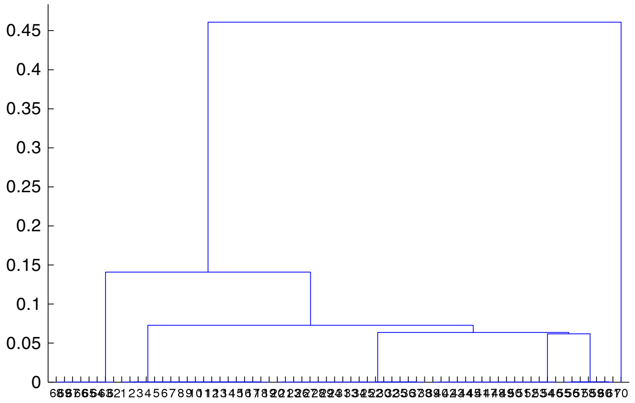

Fig. 14. The dendrogram through the proposed algorithm for data set $\mathcal{T}_1$.

The overall performance for the jth type of tea is determined as $\bar{Y}_j = (m_{\bar{Y}_j}, \alpha_{\bar{Y}_j}, \beta_{\bar{Y}_j})_T$ where

$$m_{\bar{Y}_j} = \frac{1}{4}\sum_{k=1}^{4} m_{Y_{jk}}, \qquad \alpha_{\bar{Y}_j} = \frac{1}{4}\sum_{k=1}^{4} \alpha_{Y_{jk}}, \qquad \beta_{\bar{Y}_j} = \frac{1}{4}\sum_{k=1}^{4} \beta_{Y_{jk}}.$$

Thus, we run the proposed clustering algorithm on the data set $\mathcal{T} = \{\bar{Y}_1, \ldots, \bar{Y}_{69}\}$. According to the graphical method of correlation comparisons shown in Fig. 11, $\gamma = 6$ is a suitable estimate. In the same way, we implement the SCA for

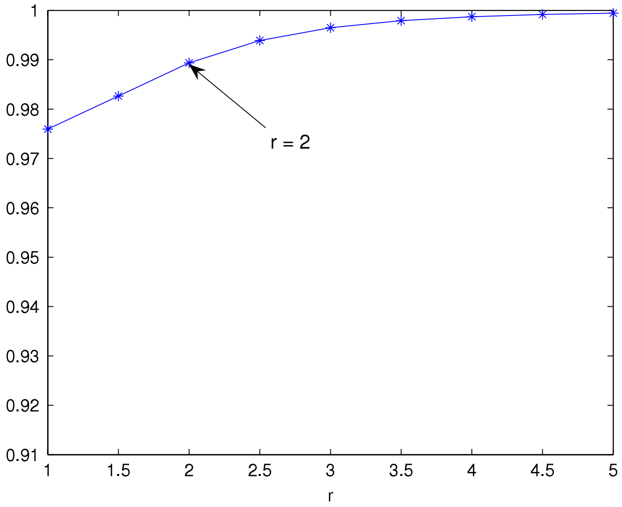

Fig. 15. The correlation coefficients between $\{\tilde{J}(X_k)_{\gamma+0.5}|k=1,\ldots,n\}$ and $\{\tilde{J}(X_k)_{\gamma+1}|k=1,\ldots,n\}$, $\gamma=0.5,1,1.5,\ldots,$ for Student data.

Table 3
Student data for Example 6.

No.	Student	Mathematics 1	Mathematics 2	Physics 1	Physics 2
1	Tom	(15, 0)	(12, 2)	(10, 10)	(15, 1)
2	David	(9, 0)	(16, 2)	(12, 2)	(10, 0)
3	Bob	(6, 0)	(10.5, 0.5)	(16.5, 3.5)	(16, 2)
4	Jane	(12, 2)	(19, 1)	(19, 0)	(11, 1)
5	Joe	(1, 1)	(8, 2)	(12, 2)	(14, 0)
6	Jack	(1, 0)	(5, 1)	(9, 0)	(7.5, 1.5)

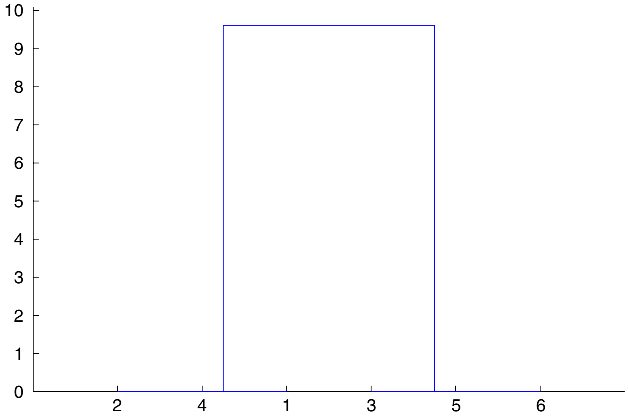

Fig. 16. The dendrogram through the proposed algorithm for Student data.

the data set and then single linkage hierarchical algorithm on the final set $\mathcal{T}$. The result is shown as the dendrogram in Fig. 12. This dendrogram clearly indicates that there are five well-separated clusters and hence the optimal cluster number $c^*=5$. The corresponding cluster centers are: $Z_1=(0.4359,0.1841,0.2480)_T$, $Z_2=(0.3754,0.1432,0.2476)_T$, $Z_3=(0.3135,0.1283,0.2499)_T$, $Z_3=(0.2517,0.1249,0.250)_T$ and $Z_5=(0.1251,0.0625,0.2500)_T$. The clustering results showed that data no. 1–19 belong to Grade 1, data numbers 20–38 belong Grade 2, data numbers 39–54 belong to Grade 3, data numbers 55–61 belong to Grade 4 and data numbers 62–69 belong to Grade 5. We mention that these clustering results are the same as Hung and Yang [9], but, in [9], the cluster number needs to be given a priori.

White-tip Oolong is the best and most famous tea in Taiwan. Experts rate its appearance, tincture, liquid color and aroma as *perfect*, *good*, *good* and *perfect*, respectively. The overall rating, for $\bar{Y}_{70}$, is $(0.875,0.25,0.125)_T$. We now add the White-tip Oolong tea $\bar{Y}_{70}$ to the data set $\mathcal{T}$ to investigate its effect on the original 69 types of tea. We also run the proposed clustering algorithm on this new data set, $\mathcal{T}_1=\{\bar{Y}_1,\ldots,\bar{Y}_{69},\bar{Y}_{70}\}$. According to the graphical method of correlation comparisons

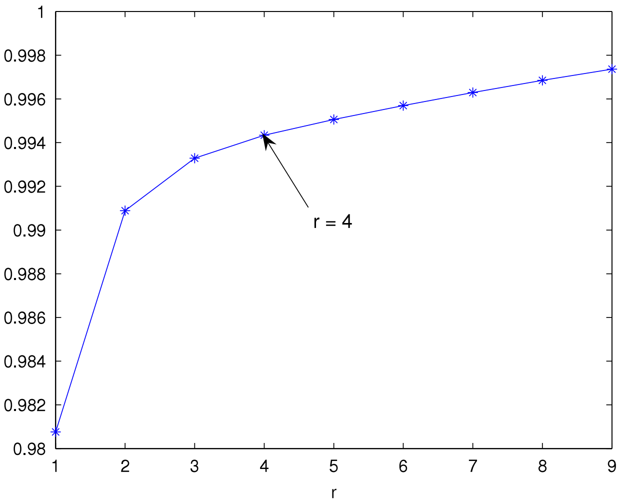

Fig. 17. The correlation coefficients between $\{\tilde{J}(X_k)_\gamma | k = 1, \dots, n\}$ and $\{\tilde{J}(X_k)_{\gamma+1} | k = 1, \dots, n\}$ for blood pressure data.

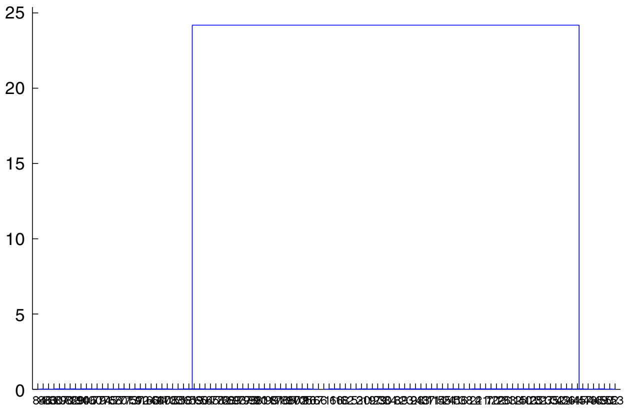

Fig. 18. The dendrogram through the proposed algorithm for blood pressure data.

shown in Fig. 13, $\gamma = 8$ is a suitable estimate, and the dendrogram is shown in Fig. 14. This dendrogram clearly indicates that there are six well-separated clusters and hence the optimal cluster number $c^* = 6$. Please note that the last cluster only includes the data number 70. In addition, the clustering results keep the same five clusters of the 69 types of tea and treat the White-tip Oolong tea as single cluster. Thus, White-tip Oolong must be regarded as a special tea according to these clustering results. This illustrates that the proposed clustering algorithm is robust and able to tolerate this special (outlier) tea.

We next consider two fuzzy data sets from [10] where the data are symmetric TFNs. That is, the left spread is equal to the right spread in which each datum is represented by (m, α).

Example 6. The student data set (cf. [16]) consists of four attributes (Mathematics 1, Mathematics 2, Physics 1 and Physics 2) and six students. D'Urso and Giordani [10] used the symmetric TFNs to model this data set. Table 3 shows the corresponding fuzzy data. According to the graphical method of correlation comparisons shown in Fig. 15, $\gamma = 2$ is a suitable estimate. We implement the SCA for the data set and then single linkage hierarchical algorithm on the final data. The dendrogram shown in Fig. 16 clearly indicates that there are two well-separated clusters and hence the optimal cluster number c^* is 2. The clustering results show that the data points 1, 2 and 4 belong to cluster 1 and the data points 3, 5 and 6 belong to cluster 2. That is, Tom, David and Jane belong to cluster 1 and Bob, Joe and Jack belong to cluster 2. The corresponding cluster centers are: $Z_1 = ((9.606, 0.381), (15.498, 1.658), (13.579, 2.461), (11.288, 0.479))$ and $Z_2 = ((3.113, 0.550), (9.004, 1.426), (13.305, 2.413), (14.061, 0.783))$. From these two cluster centers, we find that (i) the members of cluster 1 have high marks in Mathematics 2 and Physics and the moderate mark in Mathematics 1, (ii) the members of cluster 2 have very lower mark in Mathematics 1, moderate mark in Mathematics 2 and higher mark in Physics. In the sense, the students in cluster 2 have bad performance in Mathematics.

Table A
Quality comparison of 69 types of tea tree in Taiwan.

No.	Type	Appearance	Tincture	Liquid color	Aroma	$\bar{Y}$
1	Bai Mao Hou	*Perfect*	*Poor*	*Poor*	*Good*	(0.5625, 0.2500, 0.1875)
2	Hei Mao Hou	*Good*	*Poor*	*Poor*	*Good*	(0.5000, 0.2500, 0.2500)
3	Qing Xin Hei Nou	*Good*	*Poor*	*Poor*	*Good*	(0.5000, 0.2500, 0.2500)
4	Qui Zi Keng Bai Mao	*Good*	*Poor*	*Poor*	*Good*	(0.5000, 0.2500, 0.2500)
5	Da Nan Wan Bai Mao	*Perfect*	*Bad*	*Bad*	*Good*	(0.4375, 0.1250, 0.1875)
6	Qing Xin Oolong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
7	Dan Shui Qing Xin	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
8	Bu Zhi Chun	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
9	Tao Ren Chong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
10	Wan Chong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
11	Hong Xin Da Nou	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
12	Bai Xin Oolong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
13	Shui Xian	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
14	Gui Hua Chong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
15	Niu Pu Chong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
16	Lin Kou Heng Zhe	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
17	Feng Zi Lin	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
18	Pu Xin Chong	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
19	Da Hu Wei	*Good*	*Bad*	*Poor*	*Good*	(0.4375, 0.1875, 0.2500)
20	Hong Xin Oolong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
21	Fu Chou Chong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
22	Tieh Kuan Yin	*Medium*	*Bad*	*Bad*	*Perfect*	(0.3750, 0.1250, 0.1875)
23	Heng Zhe Da Ye	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
24	Gan Zi Chong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
25	San Cha Zhi Lan	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
26	Ying Zhi Hong Xin	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
27	Da Ye Oolong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
28	Tian Gong Chong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
29	Gan Zi Chong (Huang)	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
30	Wen Shen Da Ye	*Good*	*Bad*	*Poor*	*Medium*	(0.3750, 0.1875, 0.2500)
31	Hei Mian Zao Chong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
32	Qing Xin Zao Chong	*Medium*	*Bad*	*Poor*	*Good*	(0.3750, 0.1875, 0.2500)
33	Lin Kou Da Ye	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
34	Zao Chong	*Good*	*Bad*	*Bad*	*Good*	(0.3750, 0.1250, 0.2500)
35	Han Kou Chong	*Medium*	*Bad*	*Poor*	*Good*	(0.3750, 0.1875, 0.2500)
36	Niu Shi Wu	*Good*	*Bad*	*Poor*	*Medium*	(0.3750, 0.1875, 0.2500)
37	Ping Shui Chong	*Medium*	*Bad*	*Poor*	*Good*	(0.3750, 0.1875, 0.2500)
38	Yan Chuan Chong	*Good*	*Bad*	*Poor*	*Medium*	(0.3750, 0.1875, 0.2500)
39	Da Ye Zhu Ye	*Medium*	*Bad*	*Bad*	*Good*	(0.3125, 0.1250, 0.2500)
40	Tao Ren Wu	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
41	Hu Nan Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
42	Huang Zhi Chong	*Medium*	*Bad*	*Bad*	*Good*	(0.3125, 0.1250, 0.2500)
43	Ji Long Jin Gui	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
44	Wu Jin Chong	*Medium*	*Bad*	*Bad*	*Good*	(0.3125, 0.1250, 0.2500)
45	Jin Gui Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
46	Da Ji Ling Chong	*Medium*	*Bad*	*Poor*	*Medium*	(0.3125, 0.1875, 0.2500)
47	Huang Gan	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
48	Zhi Lan Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
49	Shi Tea	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
50	Bai Xin Wu Yi	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
51	Mao Er Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
52	Ji Long Bai Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
53	Zhu Ye Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
54	Yu Zhi Chong	*Good*	*Bad*	*Bad*	*Medium*	(0.3125, 0.1250, 0.2500)
55	Xiao Ye Zhu Ye	*Medium*	*Bad*	*Bad*	*Medium*	(0.2500, 0.1250, 0.2500)
56	Shen Man Chong	*Medium*	*Bad*	*Bad*	*Medium*	(0.2500, 0.1250, 0.2500)
57	Bai Chong	*Medium*	*Bad*	*Bad*	*Medium*	(0.2500, 0.1250, 0.2500)
58	Bai Ye Chong	*Medium*	*Bad*	*Bad*	*Medium*	(0.2500, 0.1250, 0.2500)
59	Yellow Tea	*Medium*	*Bad*	*Bad*	*Medium*	(0.2500, 0.1250, 0.2500)
60	Manipuri	*Good*	*Bad*	*Bad*	*Bad*	(0.1875, 0.0625, 0.2500)
61	Shan	*Good*	*Bad*	*Bad*	*Bad*	(0.1875, 0.0625, 0.2500)
62	Gao Lu Chong	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
63	Indigenou	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
64	Nan Tou Shen Tea	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
65	Japuri	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
66	A Sa Mu	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
67	Mian Dian Chong	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
68	Shan Tea	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)
69	Kyang	*Medium*	*Bad*	*Bad*	*Bad*	(0.1250, 0.0625, 0.2500)

Table B
Blood pressure data.

No.	Systolic pressure	Diastolic pressure	No.	Systolic pressure	Diastolic pressure	No.	Systolic pressure	Diastolic pressure
1	(131.0, 26.0)	(90.0, 28.0)	37	(131.5, 28.5)	(84.5, 17.5)	73	(147.0, 40.0)	(68.0, 23.0)
2	(134.5, 25.5)	(91.0, 27.0)	38	(127.5, 19.5)	(84.5, 22.5)	74	(144.0, 31.0)	(72.5, 17.5)
3	(130.0, 27.0)	(88.0, 29.0)	39	(136.0, 31.0)	(83.0, 30.0)	75	(147.5, 30.5)	(69.0, 17.0)
4	(132.5, 24.5)	(93.0, 27.0)	40	(140.0, 33.0)	(85.5, 31.5)	76	(152.0, 28.0)	(78.5, 14.5)
5	(135.5, 32.5)	(86.5, 32.5)	41	(127.0, 28.0)	(84.0, 17.0)	77	(152.5, 32.5)	(80.5, 16.5)
6	(132.5, 28.5)	(89.5, 18.5)	42	(140.0, 41.0)	(90.5, 33.5)	78	(153.0, 32.0)	(79.0, 14.0)
7	(133.5, 26.5)	(93.5, 30.5)	43	(131.0, 40.0)	(82.5, 31.5)	79	(145.5, 35.5)	(72.0, 30.0)
8	(132.0, 28.0)	(88.0, 19.0)	44	(126.5, 19.5)	(81.5, 20.5)	80	(147.5, 29.5)	(70.5, 16.5)
9	(130.0, 27.0)	(85.0, 30.0)	45	(123.0, 20.0)	(81.5, 20.5)	81	(155.5, 29.5)	(79.5, 19.5)
10	(134.5, 29.5)	(93.5, 21.5)	46	(145.0, 31.0)	(93.0, 30.0)	82	(147.0, 31.0)	(72.0, 15.0)
11	(129.0, 27.0)	(85.0, 29.0)	47	(124.0, 20.0)	(80.5, 20.5)	83	(146.5, 26.5)	(73.0, 19.0)
12	(139.0, 33.0)	(87.5, 28.5)	48	(122.5, 18.5)	(80.5, 24.5)	84	(148.0, 30.0)	(72.0, 16.0)
13	(135.0, 26.0)	(95.0, 29.0)	49	(146.0, 31.0)	(94.0, 28.0)	85	(150.5, 35.5)	(79.5, 29.5)
14	(127.5, 25.5)	(86.0, 30.0)	50	(145.5, 33.5)	(93.0, 32.0)	86	(148.5, 28.5)	(71.5, 17.5)
15	(132.5, 27.5)	(95.0, 30.0)	51	(148.5, 40.5)	(86.5, 38.0)	87	(147.0, 42.0)	(71.5, 20.5)
16	(132.5, 39.5)	(88.0, 32.0)	52	(147.5, 33.5)	(93.5, 29.5)	88	(153.0, 28.0)	(78.5, 18.5)
17	(129.0, 31.0)	(88.5, 18.5)	53	(148.0, 32.0)	(92.5, 31.5)	89	(148.5, 38.5)	(70.5, 23.5)
18	(133.5, 19.5)	(88.5, 20.5)	54	(150.0, 36.5)	(88.0, 40.0)	90	(147.0, 36.0)	(71.5, 26.5)
19	(128.0, 19.0)	(91.0, 23.0)	55	(135.5, 40.5)	(75.0, 38.0)	91	(155.0, 43.0)	(78.0, 20.0)
20	(132.0, 18.0)	(87.0, 23.0)	56	(132.5, 21.5)	(78.0, 24.5)	92	(150.0, 30.0)	(74.5, 15.5)
21	(135.0, 41.0)	(88.5, 32.5)	57	(150.0, 26.0)	(74.5, 22.0)	93	(157.0, 39.0)	(76.5, 21.5)
22	(135.0, 27.0)	(93.0, 18.0)	58	(152.5, 48.5)	(87.5, 32.5)	94	(147.5, 29.5)	(73.5, 18.5)
23	(133.5, 18.5)	(86.5, 21.5)	59	(152.0, 47.0)	(87.0, 33.0)	95	(157.0, 39.0)	(75.5, 22.5)
24	(126.0, 27.0)	(84.5, 27.5)	60	(150.5, 43.5)	(85.0, 36.0)	96	(153.0, 31.0)	(76.0, 16.0)
25	(136.5, 26.5)	(94.5, 20.5)	61	(153.5, 47.5)	(87.0, 33.0)	97	(154.0, 35.0)	(76.0, 30.0)
26	(136.0, 40.0)	(90.5, 33.5)	62	(138.0, 26.0)	(84.0, 22.0)	98	(155.0, 37.0)	(77.5, 27.5)
27	(135.5, 28.5)	(95.0, 19.0)	63	(144.0, 48.0)	(79.5, 35.5)	99	(154.0, 42.0)	(78.0, 24.0)
28	(128.5, 18.5)	(87.0, 20.0)	64	(144.5, 47.5)	(79.0, 37.0)	100	(148.0, 28.0)	(73.5, 20.5)
29	(132.0, 28.0)	(85.0, 18.0)	65	(147.5, 46.5)	(82.0, 34.0)	101	(155.5, 37.5)	(77.5, 24.5)
30	(129.5, 40.5)	(86.0, 32.0)	66	(152.0, 48.0)	(82.5, 34.5)	102	(154.5, 36.5)	(75.5, 28.5)
31	(137.0, 30.0)	(94.0, 18.0)	67	(145.5, 46.5)	(79.0, 33.0)	103	(149.5, 29.5)	(76.5, 17.5)
32	(138.0, 38.0)	(93.5, 33.5)	68	(144.0, 47.0)	(77.0, 32.0)	104	(150.0, 30.0)	(74.5, 28.5)
33	(136.0, 42.0)	(90.5, 33.5)	69	(151.5, 28.5)	(82.5, 17.5)	105	(152.5, 32.5)	(77.5, 17.5)
34	(130.5, 38.5)	(85.5, 35.5)	70	(144.5, 34.5)	(71.0, 32.0)	106	(147.0, 41.0)	(73.5, 22.5)
35	(139.0, 38.0)	(89.0, 33.0)	71	(147.0, 35.0)	(68.5, 28.5)	107	(150.0, 36.0)	(78.0, 28.0)
36	(141.5, 32.5)	(89.5, 29.5)	72	(153.5, 30.5)	(78.0, 12.0)	108	(151.5, 40.5)	(74.0, 22.0)

Example 7. The data set has 108 patients who the daily systolic and diastolic blood pressures are recorded. The data are shown in Table B of Appendix. First, the graphical method of correlation is used to find a suitable γ. From Fig. 17, the estimate of γ is 4. We implement the SCA for the data set and then single linkage hierarchical algorithm on the final data. The dendrogram as shown in Fig. 18 clearly indicates that there are two well-separated clusters and hence the optimal cluster number $c^* = 2$. The corresponding cluster centers are: $Z_1 = ((132.509, 28.046), (88.446, 25.360))$ and $Z_2 = ((150.397, 33.211), (75.201, 21.114))$. It means that cluster 1 has low values of systolic blood pressure for both of centers and spreads and high values of diastolic blood pressure for both of centers and spreads and cluster 2 has in an opposite way. Furthermore, we have the data numbers 1–50, 52–53, 56 and 62 belong to cluster 1 and the data numbers 51, 54–55, 57–61 and 63–108 belong to cluster 2. These clustering results are slightly different from [10]. But the characteristics of clusters 1 and 2 are the same as theirs.

5. Conclusions

In this paper, we proposed a robust clustering method which was presented to be robust to cluster number, initial guess and outliers and cluster shapes for clustering fuzzy data. Numerical examples actually demonstrated the effectiveness of the proposed method. In particular, we apply the proposed clustering algorithm to Taiwanese tea evaluation. The clustering results show that the algorithm can retain the original grades of the 69 types of tea (or the market price) and rank White-tip Oolong tea as an exceptional tea boosting consumer confidence and a better understanding of Taiwanese tea despite the addition of any new or special tea added to the classification process. Thus, the method can assist the Taiwan Tea Experiment Station to successfully formulate an evaluation system for tea quality. Moreover, we also applied the method to two data sets from [10]. We found that it is good for analyzing these real data sets.

In Taiwan, there has been a rapid development in higher education in the last decade. Now the numbers of university-level institutions have reached 160. To face intense competition, these institutions strive to maintain a leading position by offering quality teaching, research and service. Again this background, the Taiwan Assessment and Evaluation Association (TWAEA), a non-profit organization, was established in 2000 to provide third-party evaluation of the performance of the various universities. The function of TWAEA is to help each institution understand the strength, weakness, opportunity and threat of its subordinate departments. Therefore, it is necessary to develop more objective performance evaluation models for various academic departments. In the future, we will apply the proposed robust clustering algorithm in establishing academic performance evaluation in the higher education system in Taiwan.

Appendix

See Tables A and B.

References

[1] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[3] M.S. Yang, A survey of fuzzy clustering, Math. Comput. Modelling 18 (1993) 1–16.

[4] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods For Classification Data Analysis and Image Recognition, Wiley, New York, 1999.

[5] A. Baraldi, P. Blonda, A survey of fuzzy clustering algorithms for pattern recognition-part I and II, IEEE Trans. Syst. Man Cybern. B 29 (1999) 778–801.

[6] S.F. Schnatter, On statistical inference for fuzzy data with applications to descriptive statistics, Fuzzy Sets and Systems 50 (1992) 143–165.

[7] R.J. Hathaway, J.C. Bezdek, W. Pedrycz, A parametric model for fusing heterogeneous fuzzy data, IEEE Trans. Fuzzy Syst. 4 (1996) 270–281.

[8] M.S. Yang, C.H. Ko, On a class of fuzzy c-numbers clustering procedures for fuzzy data, Fuzzy Sets and Systems 84 (1996) 49–60.

[9] W.L. Hung, M.S. Yang, Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation, Fuzzy Sets and Systems 150 (2005) 561–577.

[10] P. D'Urso, P. Giordani, A weighted fuzzy c-means clustering model for fuzzy data, Comput. Stat. Data Anal. 50 (2006) 1496–1523.

[11] M.S. Yang, K.L. Wu, A similarity-based robust clustering method, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 434–448.

[12] H.J. Zimmermann, Fuzzy Set Theory and its Applications, Kluwer, Dordrecht, 1991.

[13] L.A. Zadeh, Similarity relations and fuzzy orderings, Inform. Sci. 3 (1971) 177–200.

[14] R.R. Yager, D.P. Filev, Approximate clustering via the mountain method, IEEE Trans. Syst. Man Cybern. 24 (1994) 1279–1284.

[15] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[16] T. Denoeux, M.H. Masson, Principal component analysis of fuzzy data using autoassociative neural networks, IEEE Trans. Fuzzy Syst. 12 (2004) 336–349.