Expert Systems with Applications 40 (2013) 929–937

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Enhanced fuzzy clustering algorithm and cluster validity index for human perception

M. Bahar Başkır [a,*], I. Burhan Türkşen [b,c,d]

[a] Department of Statistics, Faculty of Sciences, Ankara University, 06100 Tandogan-Ankara, Turkey
[b] Department of Industrial Engineering, TOBB – Economics and Technology University, Sögütözü Cad. No: 43, Sögütözü, 06560 Ankara, Turkey
[c] Director-Knowledge/Intelligence Systems Laboratory, Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, Canada M5S 3G8
[d] NSERC, Natural Science Engineering Research Council, RGPIN 7698-05, Canada

ARTICLE INFO

Keywords:
Fuzzy clustering
Cluster validity index
α-cut
Design alternative

ABSTRACT

In this study, we propose an enhanced fuzzy clustering algorithm related to α-cut interval descriptions of fuzzy numbers and a new cluster validity index, which occurs by α-cut intervals and adding two ad-hoc functions in the compactness and separability measures. As an application, we use the enhanced fuzzy clustering algorithm and its proposed validity index to rank supplier firms of a Turkish Machinery Corporation by design alternatives. In addition, the rankings of supplier firms are determined with a proposed decision measure.

1. Introduction

Fuzzy logic introduced by Zadeh (1965) is an important tool to make coherent and adequate decisions especially based on human perceptions. The perceptions of human are widely used in Critical Engineering Design (CED) studies. A perception in a CED study can be obtained by belief space and its subspaces such as knowledge and confidence (Herling et al., 1995; Ullman, Herling, & D'Ambrosio, 1997). The aim of the CED studies is to investigate how to meet a given criterion with design alternatives. The ratings of design alternatives are indicated by a quantified common belief measure, which is computed by Bayesian equation (Prasad, 1997; Ullman et al., 1997).

The well known technique of fuzzy clustering methods is Fuzzy C-Means (FCM) algorithm, which was proposed by Bezdek (1981). Numerous variations of FCM algorithm were developed for different purposes (Çelikyılmaz & Türkşen, 2008a, 2008b; Hathaway & Bezdek, 1993; Höppner & Klawonn, 2003; Pedrycz, 2004). There are many different validity indexes for choosing the number of clusters (c) and the order of fuzziness (m) based on unsupervised feature of the conventional FCM. Some of these validity indexes are Bezdek's Partition Coefficient and Partition Entropy (1974, 1976, 1981), Çelikyılmaz and Türkşen (2008a, 2008b), Fukuyama and Sugeno's index (1989), Kim and Ramakrishna's index (2005), Kung and Lin's index (2004), Tang, Sun, and Sun (2005), Xie and Beni's index (1991).

In this paper, we develop a new fuzzy clustering algorithm based on α-cut intervals of a fuzzy number to assess the critical engineering design studies. Our enhanced fuzzy clustering algorithm includes the average Euclidean distance between α-cuts of data points and their corresponding centers, which are indicated by the lower and upper limits. Furthermore, we propose a new validity index structured by α-cuts and two ad-hoc punishing functions to select the suitable cluster number and the order of fuzziness measures. The aim of using ad-hoc functions is to get the drawbacks of ratio-type Cluster Validity Indexes (CVIs) under control.

This paper is organized as follows. In Section 2, we review the conventional FCM algorithm and some validity indexes. In Section 3, we describe our enhanced fuzzy clustering technique (FCM^{α}_{new}). In Section 4, the structure of our proposed validity index (V^{α}_{new}) is given. Further we investigate the limiting behavior of V^{α}_{new} index when $c \to k$ and $m \to \infty$ to show whether our ratio-type V^{α}_{new} index has two instinct drawbacks of the ratio-type CVIs or not. In Section 5, we show the fuzzy classification results of FCM^{α}_{new} algorithm for a CED application related to design alternatives of supplier firms in a Turkish Machinery Corporation. Besides, we show the effectiveness of our V^{α}_{new} index for the CED application. Finally, the rankings of supplier firms are investigated with a proposed decision measure in this section. Conclusions are drawn in Section 6.

2. Fuzzy clustering

In real world situations, events or objects can be explained with one or more variables between multiple inputs that affect a given output of concern. Based on the similarity among observations,

* Corresponding author. Tel.: +90 312 212 6720.
E-mail address: baskir@ankara.edu.tr (M.B. Başkır).

0957-4174/$ - see front matter
http://dx.doi.org/10.1016/j.eswa.2012.05.049

fuzzy clustering algorithms are applied for an identification of potential structures presented in a given data set. One of the well known fuzzy clustering methods is the Fuzzy C-Means (FCM) algorithm. FCM algorithm was developed by Bezdek (1981) as an optimization problem given in Eq. (1).

$$\min J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^{m}(\|x_k-v_i\|)_A,$$
$$0<u_{ik}<1,\quad \forall i,k;\sum_{i=1}^{c}u_{ik}=1,\quad \forall k;0\leqslant\sum_{k=1}^{n}u_{ik}\leqslant n,\quad \forall i \tag{1}$$

In Eq. (1), $X=\{x_1,x_2,\ldots,x_n\}\subset\Re^p$ is the data set, $\Re^p$ is the feature space, n is the number of samples of data vectors; $V=[v_1,v_2,\ldots,v_c]$, ($c-$ *cluster number*, $c\in(1,n)$) is the centers vector of clusters; $U=[u_{ik}]_{cxn}$ is the membership matrix, where u_{ik} is the membership values of x_k belonging to v_i; m is the order of fuzziness. J is an objective function like the weighted within-groups sum of squared errors function. $\|\cdot\|_A A=I$ is Euclidean norm, when $A=C^{-1}$ (where C is the covariance matrix) it would be the Mahalanobis norm. The cluster centers and related membership functions, the solutions of constrained optimization problem in Eq. (1), are given in Eqs. (2) and (3), respectively.

$$v_{i,t}=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \tag{2}$$

$$u_{ik,t}=\left[\sum_{j=1}^{c}\left(\frac{\|x_k-v_{i,t-1}\|_A}{\|x_k-v_{j,t-1}\|_A}\right)^{2/(m-1)}\right]^{-1},\quad i\neq j \tag{3}$$

The well-known FCM algorithm's steps are given as follows (Türkşen, 2006):

Step 1. Choose ($c,m,iter,\ \varepsilon$)
Step 2. Find initial cluster centers.
Step 3. ITERATE
FOR $t=1$ TO *iter*
Calculate membership values in Eq. (3)
and
Calculate cluster centers in Eq. (2).
IF $t=iter$ AND $\|v_{i,t}-v_{i,t-1}\|\leqslant\varepsilon$ THEN STOP.
NEXT t

FCM algorithm has two important information; c- the number of clusters and m-the order of fuzziness. It is difficult to select suitable (c^*,m^*) pairs because of the unsupervised behavior of FCM. There are many different validity indexes for choosing the number of clusters and the order of fuzziness for fuzzy clustering algorithms. Some of these indexes are given in Table 1.

3. Enhanced fuzzy clustering algorithm

Fuzzy logic and its membership values are widely used in the subjective assessments, where a decision maker's perception has an important role. Therefore, the use of triangular fuzzy numbers is mostly prefered for decision makers' belief assessments in critical engineering design studies.

A triangular fuzzy number (TFN) can be denoted as $A=(x_{A,L},x_{A,M},x_{A,R})$. $x_{A,L}$ and $x_{A,R}$ are the lower and upper values of the support of A-TFN, respectively. $x_{A,M}$ is the mod-value of A-TFN. The membership function of A-TFN is equal to

$$\mu_A(x)=\begin{cases}0, & x\leqslant x_{A,L}\ \textit{and}\ x\geqslant x_{A,R}\\ \frac{x-x_{A,L}}{x_{A,M}-x_{A,L}}, & x_{A,L}\leqslant x\leqslant x_{A,M}\\ \frac{x_{A,R}-x}{x_{A,R}-x_{A,M}}, & x_{A,M}\leqslant x\leqslant x_{A,R}\\ 1, & x=x_{A,M}\end{cases}$$

A-TFN is the set of real numbers, which are the neighbourhoods of fuzzified $x_{A,M}$. Accordingly, it can be approximated to $x_{A,M}$ with $[x_{A,L}^{\alpha l},x_{A,R}^{\alpha l}]$ α-cut intervals for $l=1,2,\ldots,r$.

Approximation to $x_{A,M}$ with $\left[x_{A,L}^{\alpha_l},x_{A,R}^{\alpha_l}\right]$ α-cut intervals is shown in Fig. 1.

$x_{A,L}^{\alpha_l}$ and $x_{A,R}^{\alpha_l}$ shown in Fig. 1 are the lower and upper limits of lth α-cut interval of A-*TFN*, respectively. According to Kaufmann and Gupta (1985), the algebraic operations with the use of α-cuts can achieve more sensitive results than the algebraic operations of fuzzy numbers. Let $A=(x_{A,L},x_{A,M},x_{A,R})$ and $B=(y_{B,L},y_{B,M},y_{B,R})$ be two triangular fuzzy numbers and $A^\alpha=\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]$ and $B^\alpha=\left[y_{B,L}^{\alpha},y_{B,R}^{\alpha}\right]$ be the α-cuts. Some algebraic operations of A^α and B^α are as follows (Kaufmann & Gupta, 1985):

$$A^\alpha+B^\alpha=\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]+\left[y_{B,L}^{\alpha},y_{B,R}^{\alpha}\right]=\left[x_{A,L}^{\alpha}+y_{B,L}^{\alpha},x_{A,R}^{\alpha}+y_{B,R}^{\alpha}\right]$$
$$A^\alpha-B^\alpha=\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]-\left[y_{B,L}^{\alpha},y_{B,R}^{\alpha}\right]=\left[x_{A,L}^{\alpha}-y_{B,R}^{\alpha},x_{A,R}^{\alpha}-y_{B,L}^{\alpha}\right]$$
$$A^\alpha\times B^\alpha=\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]\times\left[y_{B,L}^{\alpha},y_{B,R}^{\alpha}\right]=\left[x_{A,L}^{\alpha}\times y_{B,L}^{\alpha},x_{A,R}^{\alpha}\times y_{B,R}^{\alpha}\right]$$
$$A^\alpha/B^\alpha=\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]/\left[y_{B,L}^{\alpha},y_{B,R}^{\alpha}\right]=\left[x_{A,L}^{\alpha}/y_{B,R}^{\alpha},x_{A,R}^{\alpha}/y_{B,L}^{\alpha}\right]$$
$$k>0,\quad k.A^\alpha=k.\left[x_{A,L}^{\alpha},x_{A,R}^{\alpha}\right]=\left[k.x_{A,L}^{\alpha},k.x_{A,R}^{\alpha}\right]$$

In this study, we are interested in P-unit decision makers' common belief space with its subspaces denoted as knowledge and confidence to calculate the desirability of an alternative (where P indicates the number of decision makers). The overall desirabilities based on knowledge, confidence and belief are calculated through the Fuzzy Weighted Average method proposed by Vanegas and Labib (2001a, 2001b) (for short FWA_{VL}).

Let Cr_i be the criterion and W_i represent the decision makers' common weighted factor (relative importance) of the criterion, for all $i=1,2,\ldots,n$, where n is the number of criterions. Let $K_{ij},\ C_{ij},\ \mathcal{B}_{ij}$ be the degree of a decision maker's knowledge, confidence, common belief that the jth alternative would satisfy the ith criterion ($j=1,2,\ldots,m$, where m is the number of alternatives; $i=1,2,\ldots,n$), respectively. Decision makers' common belief ($\mathcal{B}_{ij}$) is calculated by the Bayesian equation given in Eq. (4) (Prasad, 1997).

$$\mathcal{B}_{ij}=\frac{\prod_p\left[(C_{ij}\times K_{ij})+\left(\overline{C}_{ij}\times\overline{K}_{ij}\right)\right]_p}{\prod_p\left[(C_{ij}\times K_{ij})+\left(\overline{C}_{ij}\times\overline{K}_{ij}\right)\right]_p+\prod_p\left[(C_{ij}\times\overline{K}_{ij})+\left(\overline{C}_{ij}\times K_{ij}\right)\right]_p},$$
$$p=1,\ldots,P \tag{4}$$

where $\overline{C}_{ij}=1-C_{ij}$ and $\overline{K}_{ij}=1-K_{ij}$.

The lth α-cut of the overall desirability based on knowledge for jth alternative ($l=1,2,\ldots,r$, where r is the number of α-cut intervals) denoted as $K_j^{\alpha_l}=\left[K_j^{\alpha_l a},K_j^{\alpha_l b}\right]$ is calculated through the FWA_{VL} as follows:

$$K_j^{\alpha_l a}=\min_{w_i\in\left[W_i^{\alpha_l a},W_i^{\alpha_l b}\right]}\left(\frac{\sum_{i=1}^{n}K_{ij}^{\alpha_l a}w_i}{\sum_{i=1}^{n}w_i}\right)\quad\text{and}$$
$$K_j^{\alpha_l b}=\max_{w_i\in\left[W_i^{\alpha_l a},W_i^{\alpha_l b}\right]}\left(\frac{\sum_{i=1}^{n}K_{ij}^{\alpha_l b}w_i}{\sum_{i=1}^{n}w_i}\right)$$

where $\left[W_i^{\alpha_l a}=\frac{1}{P}\left(\sum_{k=1}^{P}W_{ik}^{\alpha_l a}\right),W_i^{\alpha_l b}=\frac{1}{P}\left(\sum_{k=1}^{P}W_{ik}^{\alpha_l b}\right)\right]$ is the lth α- cut of decision makers' common weighted factor. The lth α-cut of the overall desirability based on confidence (and also belief) for jth alternative denoted as $C_j^{\alpha_l}=\left[C_j^{\alpha_l a},C_j^{\alpha_l b}\right]$ (and denoted as $\mathcal{B}_j^{\alpha_l}=\left[\mathcal{B}_j^{\alpha_l a},\mathcal{B}_j^{\alpha_l b}\right]$) is calculated through the FWA_{VL} in similar way. The 'min' and 'max' operators take the minimum and maximum