Fifteenth Int. Conf. on IFSs, Burgas, 11-12 May 2011
NIFS 17 (2011), 2, 44-51

Parameter tuning in fuzzy clustering of intuitionistic fuzzy data. Part 1

N. Karthikeyini Visalakshi[1], Rangasamy Parvathi[2] and Vassia Atanassova[3]

[1] Department of Computer Science, Kongu Engineering College
Perundurai, Tamilnadu, India
Email: karthichitru@yahoo.co.in

[2] Department of Mathematics, Vellalar College for Women
Erode – 638 012, Tamilnadu, India
Email: paarvathis@rediffmail.com

[3] Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Block 105, Sofia – 1113, Bulgaria
Email: vassia.atanassova@gmail.com

Abstract: In this paper, a comparative analysis is made on Fuzzy C-Means clustering of intuitionistic fuzzy data with five different values of parameter λ. Ongoing research also focuses, in particular, on enhancing proposed clustering algorithm to produce intuitionistic fuzzy partitions.

Keywords: Clustering, Fuzzy C-Means, Intuitionistic fuzzy data.

AMS Classification: 03E72, 68T10

1 Introduction

Clustering methods seek to organize a set of objects into clusters such that objects within a given cluster have a high degree of similarity, whereas objects belonging to different clusters have a high degree of dissimilarity. These methods have been widely applied in various areas such as taxonomy, image processing, information retrieval, data mining, etc [7]. The conventional hard clustering methods represent the sharp boundaries between the clusters by specifying the membership value as 0 or 1. But, the alternate fuzzy clustering methods represent overlapping boundaries between the clusters by specifying the degree of membership between 0 and 1, so that an object can belong to one or more clusters.

The Fuzzy C-Means (FCM) [14] is one of the most popular clustering methods based on minimization of a criterion function. However, one of the greatest disadvantages of this method is its sensitivity to presence of noise and outliers in data. In recent years, many advanced technologies have been developed to store and record large quantities of data continuously. In many cases, the data may contain errors or may only be partially complete. For example, sensor networks typically create large amounts of uncertain datasets. In other cases, the data points may correspond to objects which are only vaguely specified, and are therefore considered uncertain in their representation. Similarly, surveys and imputation techniques create data which are uncertain in nature. When clustering methods are applied to these data, their uncertainty has to be considered to obtain high quality results. The presence of uncertainty changes the nature of the underlying clusters, since it affects the computation of similarity function between different data objects.

In many applications, FCM algorithm and its variants have been proven effective for clustering natural (crisp) data and representing overlapping boundaries between the clusters using the idea of fuzzy set theory [14]. But, the real-world clustering applications face additional challenge in dealing with uncertainty among data objects. This can be solved by representing data objects in terms of fuzzy membership values, before performing clustering process. In recent years, limited attention has been paid in applying FCM algorithms to deal with fuzzy data which represents crisp data in terms of membership values [15].

Intuitionistic Fuzzy Sets (IFSs) [9] are generalized fuzzy sets, which are useful in coping with the hesitancy originating from imperfect or imprecise information. Pelekis et al. [2, 13] introduced an Intuitionistic Fuzzy Representation (IFR) scheme for colour images and an Intuitionistic Fuzzy (IF) similarity measure through which a new variant of FCM algorithm is derived. But, it is not possible to use the same procedure directly for clustering numerical datasets. Karthikeyani et al. [8] proposed a new IFR scheme using parametric membership and non-membership function for clustering numerical dataset. This paper explores the significance of parameter tuning in membership and non-membership function for various domains.

The rest of this paper is organized as follows: Section 2 discusses the related works. Section 3 presents fuzzy clustering of IF data. Section 4 summarizes the experimental analysis performed with benchmark datasets. Finally, Section 5 concludes the paper.

2 Related works

There are different variants of FCM clustering in the literature [1, 4, 16, 17, 19], and each emphasizes various aspects like centroid initialization [1], number of clusters determination [4], global optimization [19], etc. This section reviews only the research works on fuzzy clustering of fuzzy data and intuitionistic fuzzy data.

P. D'Urso and P. Giordani proposed a FCM clustering model for fuzzy data, based on a weighted dissimilarity measure for comparing pairs of fuzzy data objects, composed by two distances called center distance and spread distance. They showed c-means clustering model in different information paradigm. LR fuzzy data model is used to represent the general class of fuzzy data [15].

Yang and Ko [12] derived new types of fuzzy clustering procedures in dealing with fuzzy data called as Fuzzy C-Numbers (FCN) clusterings. They have constructed these FCNs for LR-type, triangular, trapezoidal and normal fuzzy numbers.

An Alternative Fuzzy *C*-Numbers (AFCN) clustering algorithm is presented in [21], for *LR*-type fuzzy numbers based on an exponential-type distance function. On the basis of the gross error sensitivity and influence function, this exponential-type distance is claimed to be robust with respect to noise and outliers. The AFCN clustering algorithm is more robust than the FCN clustering algorithm presented by Yang and Ko [11].

In [20], the fuzzy clustering based on intuitionistic fuzzy relation is discussed. The clustering algorithm uses similarity-relation matrix, obtained by n-step procedure based on max-t and min-s compositions.

In [18], Vicenc Torra et al. introduced a method to define intuitionistic fuzzy partitions from the result of different fuzzy clustering algorithms such as FCM, entropy based FCM and FCM with tolerance. In this approach, the intuitionistic fuzzy partition permits to cope with the uncertainty present in the execution of different fuzzy clustering algorithms with the same data and with the same parameterization.

In [22], Z. Xu. et al have developed a straightforward and practical algorithm for clustering IFS, which consists of the following two steps: Firstly, it employs the derived association coefficients of IFS to construct an association matrix, and utilizes a procedure to transform it into an equivalent association matrix. Secondly, it constructs the α-cutting matrix of the equivalent association matrix, and then classifies the IFS under the given confidence levels.

In [2, 13], N. Pelekis et al. investigated the issue of clustering intuitionistic fuzzy representation of images. To achieve that they proposed a clustering approach based on the FCM algorithm utilizing a novel similarity metric defined over IFS. The performance of the modified FCM algorithm is evaluated for object clustering in the presence of noise and image segmentation. It is proved that clustering intuitionistic fuzzy image representations is more effective, noise tolerant and efficient as compared with the conventional FCM clustering of both crisp and fuzzy image representations.

3 Fuzzy clustering of intuitionistic fuzzy data objects

3.1 Intuitionistic fuzzy sets

Fuzzy sets are designed to represent or manipulate data and information possessing non-statistical uncertainties. Since Zadeh [10] introduced the concept of fuzzy sets, various notions of high-order fuzzy sets have been proposed. Among them, IFSs introduced by Atanassov [9], have captured the attention of many researchers in the last decades. This is mainly due to the fact that IFSs are consistent with human behaviour, by reflecting and modelling the hesitancy present in real-life situations. Since IFSs can present the degrees of membership and non-membership with a degree of hesitancy, the knowledge and semantic representation become more meaningful and applicable. An intuitionistic fuzzy set is defined as a generalization of a fuzzy set.

Definition 3.1. Let a set E be fixed. A fuzzy set on E is an object $\overline{A}$ of the form

$$\overline{A} = \left\{ \left\langle x, \mu_{\overline{A}}(x) \right\rangle \middle| \, x \in E \right\} \tag{1}$$

where $\mu_{\overline{A}} : E \to [0,1]$ defines the degree of membership of the element $x \in E$ to the set $\overline{A} \subset E$. For every element $x \in E$, $0 \le \mu_{\overline{A}}(x) \le 1$.

Definition 3.2. An IFS A is an object of the form

$$A = \left\{ \left\langle x, \mu_A(x), \nu_A(x) \right\rangle \middle| x \in E \right\} \tag{2}$$

where $\mu_A : E \to [0,1]$ and $\nu_A : E \to [0,1]$ define the degree of membership and non-membership, respectively, of the element $x \in E$ to the set $A \subset E$. For every element $x \in E$, it holds that $0 \le \mu_A(x) + \nu_A(x) \le 1$.

For every $x \in E$, if $\nu_A(x) = 1 - \mu_A(x)$, then A represents a fuzzy set. The function

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \tag{3}$$

represents the degree of hesitancy of the element $x \in E$ to the set $A \subset E$.

3.2 Intuitionistic fuzzy representation of numerical data objects

The proposed IFDFC algorithm requires that each element is to be converted into a pair of membership and non membership values. A new procedure for intuitionistic fuzzy representation of numeric data is derived, by modifying the definition for intuitionistic fuzzy representation of digital image [6]. In this process, the crisp dataset is first transferred to fuzzy domain and sequentially into the intuitionistic fuzzy domain, where the clustering is performed.

Let *X* be the dataset of *n* objects and each object contains *d* features. The proposed IF data clustering requires that each data element x_{ij} belongs to IFS X' by a degree $\mu_i(x_j)$ and does not belong to X' by a degree $\nu_i(x_j)$, where *i* and *j* represent objects and features of the dataset, respectively.

A membership function $\overline{\mu_i}(x_j)$ for intermediate fuzzy representation is defined by

$$\overline{\mu_i}(x_j)=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)}$$

where $i=1,2,...,n$ and $j=1,2,...,d$.

The intuitionistic fuzzification based on the family of parametric membership and non-membership function, used for clustering, is defined respectively by

$$\mu_i(x_j;\lambda)=1-(1-\overline{\mu_i}(x_j))^{\lambda}$$

and

$$\nu_i(x_j;\lambda)=(1-\overline{\mu_i}(x_j))^{\lambda(\lambda+1)}$$

where $\lambda\in[0,1]$.

The intuitionistic fuzzification converts crisp dataset $X(x_{ij})$ into intuitionistic fuzzy dataset $X'(x_{ij},\mu_i(x_j),\nu_i(x_j))$.

3.3 Modified fuzzy C-means clustering

In [13], Pelekis proposed a new variant of FCM clustering algorithm that copes with uncertainty and a similarity measure between intuitionistic fuzzy sets, based on the membership and non membership values of their elements. The procedure used in modified FCM is same as conventional FCM, except in similarity measure used to compute the membership degree of the object to cluster. Instead of Euclidean distance in conventional FCM, the modified FCM applies IF similarity measure for any two elements namely *A* and *B* as follows:

$$S_1(A,B)=\frac{S'(\mu_A(x_i),\mu_B(x_i))+S'(\nu_A(x_i),\nu_B(x_i))}{2} \tag{4}$$

where

$$S'(A',B')=\begin{cases}\dfrac{\sum_{i=1}^{n}\min(A'(x_i),B'(x_i))}{\sum_{i=1}^{n}\max(A'(x_i),B'(x_i))}, & A'\cup B'\neq\Phi\\ 1, & A'\cup B'=\Phi\end{cases} \tag{5}$$

The modified FCM algorithm is described in Fig. 1. Initially, *C* numbers of centroids are randomly selected from the intuitionistic fuzzy objects, which contain both membership and non-membership values. Next, the membership degree of each object to each cluster U_{ij} is computed using IF similarity measure as in equation (6). The centroids are then updated using Cluster Membership Matrix (CMM) U_{ij} and corresponding membership and non-membership degrees of centroids V_i are also computed. The above two steps are repeated, until it reaches convergence.

Algorithm. Modified FCM

Input: Dataset of *n* objects with *d* features, value of *C* and fuzzification value $m > 1$
Output: Cluster Membership Matrix U_{ij} for n objects and *C* clusters
Procedure:
Step 1: Determine initial centroids by selecting *c* random intuitionistic fuzzy objects.
Step 2: Compute the values for CMM represented by U_{ij}, using

$$\underset{\substack{1\le i\le c\\ 1\le j\le n}}{\forall} U_{ij} = \begin{cases} \dfrac{\left(S_1(x_j - V_i)\right)^{\frac{1}{1-m}}}{\sum_{l=1}^{c}\left(S_1(x_j - V_l)\right)^{\frac{1}{1-m}}}, & I_j = \phi \\ \begin{cases} 0, & i \notin I_j \\ \sum_{i\in I_j} U_{ij} = 1, & i \in I_j,\ I_j \neq \phi \end{cases} \end{cases} \tag{6}$$

where

$$\underset{\forall 1\le j\le n}{I_j} = \left\{ i \,\middle|\, 1 \le i \le c; S_1(x_j, V_i) = 0 \right\}$$

Step 3: Update the centroids' matrix V_i using

$$\underset{1\le i\le c}{\forall} V_i = \frac{\sum_{j=1}^{n} (U_{ij})^m x_j}{\sum_{j=1}^{n} (U_{ij})^m} \tag{7}$$

Step 4: Compute membership and non-membership degrees of V_i
Step 5: Repeat step 2 to step 4 until converges.

Fig. 1. Modified Fuzzy C-Means Algorithm

4 Experimental analysis

The main purpose of this work is to explore the role of intuitionistic fuzzification of numerical data and significance of parameter tuning in the process of FCM clustering. The experimental analysis is performed with ten benchmark datasets available in the UCI machine learning data repository [11]. The results of FCM clustering of IF data is analyzed with different values of λ, in the computation of membership and non-membership values. The information about the datasets is shown in Table 1.

The performance of the clustering algorithm is measured in terms of two external validity measures [3, 5] namely F-Measure and Entropy. The external validity measures test the quality of clusters by comparing the results of clustering with the "ground truth" (true class labels). Both these measures have a value between 0 and 1. In case of F-measure, the value 1 indicates that the data clusters are exactly same. But, the value 1 signifies that the data clusters are entirely different for Entropy measure.

Table 2 and Table 3 depict the performance on FCM clustering of IF Data with five different values of λ, in terms of F-Measure and Entropy respectively. From the results of these tables, it identified that the assignment value to parameter λ depends on the characteristics of dataset.

Table 1. Details of datasets

S. No.	Dataset	No. of Attributes	No. of Classes	No. of Instances
1	Australian	14	2	690
2	Breast Cancer	10	2	699
3	Dermatology	34	6	366
4	Image Segmentation	19	7	2310
5	Mammography	5	2	961
6	Pima Indian Diabetes	8	2	768
7	Satellite Image	36	7	6435
8	Spam E-Mail	58	2	4601
9	Thyroid	21	3	7200
10	White Wine	11	7	4898

Table 2. Performance of FCM Clustering based on F-Measure

Dataset	λ = **0.95**	λ = **0.85**	λ = **0.75**	λ = **0.65**	λ = **0.55**
Australian	0.465	0.467	0.468	**0.521**	0.468
Breast Cancer	0.662	0.662	**0.701**	0.654	0.621
Dermatology	**0.819**	0.754	0.764	0.754	0.734
Image Segmentation	0.683	**0.724**	0.693	0.664	0.684
Mammography	0.782	0.780	0.741	0.788	**0.852**
Pima Indian Diabetes	**0.521**	0.482	0.421	0.400	0.399
Satellite Image	0.641	0.634	0.624	0.618	**0.692**
Spam E-Mail	0.436	0.436	**0.449**	0.400	0.401
Thyroid	0.615	0.641	0.625	0.646	**0.648**
White Wine	**0.375**	0.362	0.353	0.334	0.325

Table 3. Performance of FCM Clustering based on Entropy

Dataset	λ = **0.95**	λ = **0.85**	λ = **0.75**	λ = **0.65**	λ = **0.55**
Australian	0.217	0.235	0.249	**0.205**	0.286
Breast Cancer	0.079	0.087	**0.069**	0.088	0.085
Dermatology	**0.335**	0.374	0.356	0.379	0.406
Image Segmentation	0.719	**0.633**	0.690	0.728	0.652
Mammography	0.503	0.507	0.505	0.510	**0.492**
Pima Indian Diabetes	**0.312**	0.346	0.335	0.332	0.315
Satellite Image	0.779	0.772	0.782	0.784	**0.683**
Spam E-Mail	0.310	0.325	**0.292**	0.312	0.314
Thyroid	0.272	**0.252**	0.300	0.298	0.294
White Wine	**0.712**	0.801	0.811	0.792	0.776

5 Conclusion

In this paper, a comparative analysis is made on FCM clustering of IF data with five different values of parameter λ. It can be concluded that it is important to assign appropriate value to parameter λ, according to the nature of dataset. In future, applying optimization algorithm for tuning of parameter λ will help in producing superior quality clusters. Ongoing research also focuses, in particular, on enhancing proposed clustering algorithm to produce intuitionistic fuzzy partitions.

Acknowledgements

The authors would like to thank the Department of Science and Technology, New Delhi, India and Ministry of Education and Science, Sofia, Bulgaria, for their financial support to the Bilateral Scientific Cooperation Research Programme INT/Bulgaria/B-2/08 and BIn-02/09.

References

[1] Dae-Won Kim, Kwang Hyung Lee, Doheon Lee (2004) A novel initialization scheme for the fuzzy c-means algorithm for color clustering. Pattern Recognition Letters 25(2): 227–237.

[2] Dimitrios K. Iakovidis, Nikos Pelekis, Evangelos E. Kotsifakos, Ioannis Kopanakis (2008) Intuitionistic fuzzy clustering with applications in computer vision. In: Advanced concepts for intelligent vision system. Springer, Berlin.

[3] Halkidi, M., Y. Batistakis, M. Vazirgiannis. (2002) Cluster validity methods: part I. ACM SIGMOD Record. 31(2):19–27.

[4] Haojun Sun, Shengrui Wang, Qingshan Jiang (2004) FCM-based model selection algorithms for determining the number of clusters. Pattern Recognition 37(10): 2027–2037.

[5] Hui Xiong, Junjie Wu, Jian Chen (2006) K-means clustering versus validation measures: a data distribution perspective. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining**.** 779–784.

[6] Vlachos, I., G. D. Sergiadis (2007) The Role of Entropy in Intuitionistic Fuzzy Contrast Enhancement. Foundations of fuzzy logic and soft computing, Springer, Berlin.

[7] Jain A K, Murthy M N, Flynn P J (1999) Data Clustering: A Review. ACM Computing Surveys 31(3): 265–323.

[8] Karthikeyani Visalakshi, K. Thangavel, R. Parvathi (2010) An intuitionistic fuzzy approach to fuzzy clustering of numerical dataset , International Journal of Computer Theory and Engineering, 2(2), 1793–8201, 295–302.

[9] Atanassov, K. (2003) Intuitionistic fuzzy sets: past, present and future. In: Proceedings of the 3rd Conference of the European Society for Fuzzy Logic and Technology. 12–19.

[10] Lotfi A. Zadeh (1965) Fuzzy sets. Information and Control 8(3): 338–353.

[11] Merz C J, Murphy P M (1998) UCI Repository of Machine Learning Databases. Irvine, University of California, `http://www.ics.uci.eedu/~mlearn/`.

[12] Miin-Shen Yang, Cheng-Hsiu Ko (1996), On a class of fuzzy c-numbers clustering procedures for fuzzy data. Fuzzy Sets and Systems, 84:49–60.

[13] Nikos Pelekis, Dimitrios K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis (2008) Fuzzy clustering of intuitionistic fuzzy data. International journal of business intelligence and data mining. 3(1): 45–65.

[14] Pang-Ning Tan, Steinbach M., Kumar V.(2006) Cluster Analysis: Basic Concepts and algorithms. In: Introduction to Data Mining, Pearson Addison Wesley, Boston.

[15] Pierpaolo D'Urso, Paolo Giordani (2006) A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics & Data Analysis 50(6): 1496–1523.

[16] Sueli A. Mingoti, Joab O. Lima (2006) Comparing SOM neural network with fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research 174(3): 1742–1759.

[17] Tai Wai Cheng, Dimitry B.Goldgof, Lawrence O Hall (1995) Fast Clustering with application to fuzzy rule generation. In: Proceedings of the 4th IEEE International conference on fuzzy systems. 2289–2295.

[18] Torra, V. Miyamoto, S., Endo, Y. Domingo-Ferrer, J. (2008) On intuitionistic fuzzy clustering for its application to privacy. In: Proceedings of IEEE International conference on fuzzy systems. 1042–1048.

[19] Weina, Wang, Yunjie Zhang, Yi Li, Xiaona Zhang (2006) The Global Fuzzy C-Means Clustering Algorithm. In: Proceedings of the 6th World Congress on Intelligent Control and Automation. 3604–3607.

[20] Wen-Liang Hung, Jinn-Shing Lee, Cheng-Der Fuh (2004) Fuzzy Clustering Based On Intuitionistic Fuzzy Relations. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12(4): 513–530.

[21] Wen-Liang Hung Miin-Shen Yang (2005) Fuzzy clustering on *LR*-type fuzzy numbers with an application in Taiwanese tea evaluation. Fuzzy Sets and Systems 150(3): 561–577.

[22] Zeshui Xu, Jian Chen, Junjie Wu (2008) Clustering algorithm for intuitionistic fuzzy sets. Information Sciences 178(19): 3775–3790.