An Intuitionistic Fuzzy Approach to Fuzzy Clustering of Numerical Dataset

N. Karthikeyani Visalakshi, S. Parvathavarthini and K. Thangavel

Abstract Fuzzy c-means (FCM) clustering is one of the most widely used fuzzy clustering algorithms. However, the main disadvantage of this algorithm is its sensitivity to noise and outliers. Intuitionistic fuzzy set is a suitable tool to cope with imperfectly defined facts and data, as well as with imprecise knowledge. So far, there exists a little investigation on FCM algorithm for clustering intuitionistic fuzzy data. This paper focuses mainly on two aspects. Firstly, it proposes an intuitionistic fuzzy representation (IFR) scheme for numerical dataset and applies the modified FCM clustering for clustering intuitionistic fuzzy (IF) data and comparing results with that of crisp and fuzzy data. Secondly, in clustering of IF data, different IF similarity measures are studied and a comparative analysis is carried out on the results. The experiments are conducted for numerical datasets of UCI machine learning data repository.

Keywords Clustering · Fuzzy c-means · Intuitionistic fuzzy data · Intuitionistic fuzzy similarity measure

N. Karthikeyani Visalakshi (✉) · S. Parvathavarthini
Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India
e-mail: karthichitru@yahoo.co.in

S. Parvathavarthini
e-mail: varthinis@gmail.com

K. Thangavel
Periyar University, Salem, Tamil Nadu, India
e-mail: drktvelu@yahoo.com

G. S. S. Krishnan et al. (eds.), *Computational Intelligence, Cyber Security and Computational Models*, Advances in Intelligent Systems and Computing 246,
DOI: 10.1007/978-81-322-1680-3_9, © Springer India 2014

1 Introduction

Clustering algorithms 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. Clusters can be hard or fuzzy in nature based on whether each data object has to be assigned exclusively to one cluster or allowing each object to be assigned to every cluster with an associated membership value.

The Fuzzy C-Means (FCM) algorithm is sensitive to the presence of noise and outliers in data [1]. To enhance robustness of FCM, different researchers proposed different methodologies [1–3]. Intuitionistic fuzzy sets (IFSs) [4] are generalized fuzzy sets, which use the hesitancy originating from imprecise information. Pelekis et al. [5] introduced an Intuitionistic Fuzzy Representation (IFR) scheme for color images and an Intuitionistic Fuzzy (IF) similarity measure through which a new variant of FCM algorithm is derived. But this cannot be directly used for clustering numerical datasets. Hence, robust fuzzy clustering is proposed in this paper to make FCM algorithm as noise insensitive, by dealing with IF data. Real data are converted into IFR, before clustering, in order to achieve the benefit of IFSs in fuzzy clustering. A comparative study is made on fuzzy clustering of crisp, fuzzy, and IF data, and the performance of IF clustering is measured using four different IF similarity measures.

The rest of this paper is organized as follows: Sect. 2 provides discussions on IFS and IF similarity measures. Section 3 reviews the related works. The proposed method of clustering numerical dataset is described in Sect. 4. Section 5 summarizes the experimental analysis performed with benchmark datasets. Section 6 concludes the paper.

2 Background

2.1 *Intuitionistic Fuzzy Sets*

Fuzzy sets are designed to manipulate data and information possessing non-statistical uncertainties. Since Zadeh [6] introduced the concept of fuzzy sets, various notions of high-order fuzzy sets have been proposed. Among them, IFSs, introduced by Atanassov [4], can present the degrees of membership and non-membership with a degree of hesitancy.

Definition 2.1 An IFS A is an object of the form:

$$A = \{\langle x, \mu_A(x), \nu_A(x)\rangle | x \in E\} \tag{1}$$

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 $0 \leq \mu_A(x) + v_A(x) \leq 1$. If A represents a fuzzy set, for every $x \in E$, if $v_A(x) = 1 - \mu_A(x)$ and

$$\pi_A(x) = 1 - \mu_A(x) - v_A(x) \tag{2}$$

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

2.2 Intuitionistic Fuzzy Similarity Measures

Similarity measure determines the degree of similarity between two objects. Many of them are proposed by different researchers [5, 7, 8] and are applied in a wide range of applications. In the following, four IFS similarity measures used in this work for comparative analysis are reviewed.

Pelekis [5] proposed a similarity measure S_1between the IFSs A and B as

$$S_1(A,B) = \frac{S'(\mu_A(x_i), \mu_B(x_i)) + S'(v_A(x_i), v_B(x_i))}{2} \tag{3}$$

$$\text{where } S'(A', B') = \begin{cases} \frac{\sum_{i=1}^{n} \min(A'(x_i), B'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), B'(x_i))}, & A' \cup B' \neq \Phi \\ 1, & A' \cup B' = \Phi \end{cases} \tag{4}$$

where Φ is a fuzzy set for which the membership function is zero for all elements. This measure uses the aggregation of the minimum and maximum membership values in combination with those of the non-membership values.

Hung and Yang [8] extended some similarity measures of FS to IFSs,

$$S_2(A,B) = \frac{\sum_{i=1}^{n} (\min(\mu_A(x_i), \mu_B(x_i)) + \min(v_A(x_i), v_B(x_i)))}{\sum_{i=1}^{n} (\max(\mu_A(x_i), \mu_B(x_i)) + \max(v_A(x_i), v_B(x_i)))} \tag{5}$$

It focuses on the ratio of the aggregation of minimum of membership and non-membership values to the aggregation of maximum of membership and non-membership values. They also proposed a new similarity measure S_3 as in Eq. (6), which adopts exponential operation to the Hamming distance between IFSs A and B.

$$S_3(A,B) = 1 - \frac{1 - \exp\left(-\frac{1}{2}\sum_{i=1}^{n} |\mu_A(x_i) - \mu_B(x_i)| + |v_A(x_i) - v_B(x_i)|\right.}{1 - \exp(-n)} \tag{6}$$

The similarity measure S_4 as in Eq. (7) considers the hesitancy values of IFSs in computing similarity between IFSs A and B, based on normalized Hamming distance [7].

$$S_4(A,B) = \frac{1}{2n}\sum_{i=1}^{n} |\mu_A(x_i) - \mu_B(x_i)| + |v_A(x_i) - v_B(x_i)| + |\pi_A(x_i) - \pi_B(x_i)| \tag{7}$$

3 Related Works

There are different variants of FCM clustering in the literature.

D'Urso and Giordani [9] proposed a FCM clustering model for LR-type fuzzy data, based on a weighted dissimilarity measure for comparing fuzzy data objects, using center distance and spread distance. Leski [1] introduced a new ε-insensitive Fuzzy C-Means (εFCM) clustering algorithm in order to make FCM as noise insensitive. In [10], the fuzzy clustering based on IF relation is discussed. The clustering algorithm uses similarity-relation matrix, obtained by n-step procedure based on max-t and min-s compositions.

Bannerji et al. [2] proposed robust fuzzy clustering methodologies to deal with noise and outliers, by means of mega-cluster concept and robust error estimator. In [5, 11], Pelekis et al. clustered IF representation of images and proposed a clustering approach based on the FCM using a novel similarity metric defined over IFSs, which is more noise tolerant and efficient as compared with the conventional FCM clustering of both crisp and fuzzy image representations.

4 Intuitionistic Fuzzy Approach to Fuzzy Clustering

The proposed methodology for the fuzzy clustering using IFSs involves two stages, viz., intuitionistic fuzzification to convert the real scalar values into IF values and using modified FCM algorithm based on IF similarity measure to cluster IF data. Additionally, four IF similarity measures are also used for comparative analysis.

4.1 Intuitionistic Fuzzification

Following [12], a new procedure for intuitionistic fuzzification of numerical dataset is derived where the crisp dataset is first transferred to fuzzy domain and sequentially into the IF 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 an IFS X' by a degree $\mu_i(x_j)$ and does not belong to X' by a degree $v_i(x_j)$, 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

$$\overline{\mu_i}(x_j) = \frac{x_{\mathrm{ij}} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad \text{where} \quad i = 1, 2, \ldots, N \,\text{and}\, j = 1, 2, \ldots, d \tag{8}$$

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

$$\mu_i(x_j;\,\lambda) = 1 - (1 - \overline{\mu_i}(x_j))^{\lambda} \tag{9}$$

and

$$v_i(x_j;\,\lambda) = (1 - \overline{\mu_i}(x_j))^{\lambda(\lambda+1)} \quad \text{where} \quad \lambda \in [0,1] \tag{10}$$

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

4.2 Fuzzy Clustering of IF Data

In this stage, Pelekis's modified FCM [5] is applied to cluster IF data. Instead of euclidean distance in conventional FCM, the modified FCM applies IF similarity measure. The modified FCM algorithm is as follows:

Step 1. Determine initial centroids by selecting c random IF objects.
Step 2. Calculate the membership matrix U_{ij}, using

$$\underset{\substack{1 \le i \le c \\ 1 \le j \le N}}{\forall} \quad U_{\mathrm{ij}} = \begin{cases} \dfrac{\left(S_1(x_j - C_i)\right)^{\frac{1}{1-m}}}{\sum_{l=1}^{c}\left(S_1(x_j - C_l)\right)^{\frac{1}{1-m}}}, & I_j = \phi \\ \begin{cases} 0, & i \notin I_j \\ \sum_{i \in I_j} U_{\mathrm{ij}} = 1, & i \in I_j,\; I_j \neq \phi \end{cases} \end{cases} \tag{11}$$

$$\text{where} \quad \underset{\forall 1 \le j \le N}{I_j} = \left\{ i \middle| 1 \le i \le c;\, S_1(x_j, C_i) = 0 \right\}$$

Step 3. Update the centroids' matrix C_i using

$$\underset{1 \le i \le c}{\forall} C_i = \frac{\sum_{j=1}^{n} (U_{\mathrm{ij}})^m x_j}{\sum_{j=1}^{n} (U_{\mathrm{ij}})^m} \tag{12}$$

Step 4. Compute membership and non-membership degrees of C_i

Step 5. Repeat step 2 to step 4 until converges.

Initially, c number of centroids are randomly selected from the IF objects, which contain both membership and non-membership values. Next, the membership degree of each object to each cluster $\mathrm{U_{ij}}$ is computed using IF similarity measure as in Eq. (11). The centroids are then updated using cluster membership matrix, and corresponding membership and non-membership degrees of centroids C_i are also computed. Repeat the above two steps until convergence.

5 Experimental Analysis

This work explores the role of intuitionistic fuzzification of numerical data and IF similarity measures in the process of FCM clustering. The experimental analysis is carried out with five benchmark datasets in two aspects. First, the results of FCM clustering on crisp, fuzzy, and IF data are compared, and the fuzzification is done using Eq. (8). The λ value is set as 0.95, for the computation of membership and non-membership values using Eqs. (9) and (10). Next, the performance of four different IF similarity measures is evaluated.

Experiments are conducted using the breast cancer, dermatology, image segmentation, satellite image, and wine datasets available in the UCI machine learning data repository [13]. The conventional FCM algorithm is used to cluster crisp and fuzzy data, and the modified FCM using IF similarity measure is used to cluster IF data. Experiments are run 50 times on each dataset, and average values are taken for evaluation.

5.1 Cluster Evaluation Criteria

Here, the performance of fuzzy clustering algorithm is measured in terms of two external validity measures, [12, 14] the Rand index, F-measure and two fuzzy internal validity measures, fuzzy DB (FDB) index, and Xie–Beni (XB) index. The maximum value indicates good performance for Rand index and F-measure. The minimum value indicates the better performance for both FDB and XB indices.

5.2 Comparative Analysis on Crisp, Fuzzy, and IF Data

Two sets of experiments are conducted to evaluate the performance of IF representation.

5.2.1 Hard Cluster Evaluation

The first experiment compares the efficiency using hard cluster validity measures. Table 1 depicts the performance of FCM clustering on crisp, fuzzy, and IF data. It is observed that the number of iterations required for FCM clustering is highly

Table 1 Comparative analysis based on Rand index, F-measure, and number of iterations

S. No.	Dataset	Number of iterations			Rand index			F-measure		
		Crisp data	Fuzzy data	IF data	Crisp data	Fuzzy data	IF data	Crisp data	Fuzzy data	IF data
1	Breast cancer	43	17	**5**	0.521	**0.928**	0.907	0.615	0.653	**0.664**
2	Dermatology	46	12	**6**	0.701	0.672	**0.910**	0.307	0.642	**0.823**
3	Image segmentation	100	97	**14**	0.818	0.635	**0.881**	0.488	0.459	**0.691**
4	Satellite image	100	101	**34**	0.848	0.594	**0.853**	0.673	0.549	**0.713**
5	Wine	55	22	**5**	0.717	0.683	**0.911**	0.674	0.749	**0.934**

reduced, when IF data are used in clustering for all datasets. With Rand index, the performance of clustering IF data dominates that of clustering crisp data, for all datasets, and is outstanding for dermatology and wine datasets. With F-measure, the performance improvement of IF data is higher for all datasets. The F-measure is highly appreciable for wine and dermatology datasets with IF data.

5.2.2 Soft Cluster Evaluation

The second experiment compares the efficiency using fuzzy cluster validity measures. Table 2 depicts the performance of modified FCM clustering on crisp, fuzzy, and IF data in terms of FDB index and XB index. It is proved that performance of clustering IF data is better than other two approaches, for all datasets.

From the analysis, it is observed that the representation of IF data before clustering is more suitable for all numerical datasets. However, satellite image and image segmentation datasets yield better results for fuzzy data representations.

Table 2 Comparative analysis based on FDB index and XB index

S. No.	Dataset	FDB index			XB index		
		Crisp data	Fuzzy data	IF data	Crisp data	Fuzzy data	IF data
1	Breast cancer	0.525	0.468	**0.389**	0.276	0.489	**0.074**
2	Dermatology	1.256	1.115	**1.002**	0.349	0.298	**0.265**
3	Image segmentation	0.898	0.813	**0.761**	1.125	0.265	**0.722**
4	Satellite image	1.562	1.455	**1.232**	0.704	0.381	**0.671**
5	Wine	0.501	0.478	**0.312**	0.126	0.117	**0.101**

Table 3 Comparative analysis of IF similarity measures

S. No.	Dataset	Validity measure	S_1	S_2	S_3	S_4
1	Breast cancer	Rand index	0.907	0.626	0.626	0.630
		F-measure	0.664	0.763	0.763	0.762
2	Dermatology	Rand index	0.910	0.817	0.821	0.201
		F-measure	0.823	0.693	0.712	0.469
3	Image segmentation	Rand index	0.881	0.875	0.872	0.879
		F-measure	0.691	0.678	0.669	0.686
4	Satellite image	Rand index	0.854	0.828	0.831	0.185
		F-measure	0.713	0.665	0.669	0.381
5	Wine	Rand index	0.911	0.872	0.872	0.342
		F-measure	0.934	0.900	0.900	0.570

5.3 Comparative Analysis on Intuitionistic Fuzzy Similarity Measures

This experiment compares the quality of clusters obtained by the modified FCM with IF similarity measure S_1, S_2, S_3, and S_4. Table 3 shows the effect of four similarity measures based on Rand index and F-measure. From the results of the Table 3, it is identified that the similarity measure S_1 is more suitable than the other three measures for all datasets. The impacts of all four similarity measures are almost same, for image segmentation datasets.

6 Conclusion

In this paper, a novel procedure for intuitionistic fuzzification of numerical dataset is proposed and the IF data are applied to the modified FCM clustering algorithm to obtain fuzzy clusters. Experiments are conducted to study the impact of using IF data representation and IF similarity measures in FCM clustering. It can be concluded that the conversion of crisp data into IF data before clustering leads to obtain better quality clusters. It is observed that the IF similarity measure S_1 may be suitable for achieving competent fuzzy clusters. In future, applying optimization algorithm for tuning of parameter λ will help in producing superior quality clusters. Proposed algorithm may be enhanced to produce IF partitions.

References

1. J. Leski, Towards a robust fuzzy clustering. Fuzzy Sets and Systems. 137(2) (2003) 215-233.
2. Banerjee A, Dave R.N, The fuzzy mega-cluster: Robustifying FCM by Scaling down memberships. In: Lecture Notes in Artificial Intelligence, Springer (2005).

3. Bohdan S. Butkiewicz, Robust fuzzy clustering with fuzzy data. In: Advances in web intelligence, Springer, Berlin, 2005.
4. KT. Atanassov, Intuitionistic fuzzy sets: past, present and future. In: Proceedings of the 3rd Conference of the European Society for Fuzzy Logic and Technology, 2003, pp. 12-19.
5. Nikos Pelekis, Dimitrios K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis, Fuzzy clustering of intuitionistic fuzzy data. International Journal of Business Intelligence and Data Mining 3(1) (2008) 45-65.
6. L.A. Zadeh, Fuzzy sets. Information and Control 8(3) (1965) 338-353.
7. Szmidt E, Kacprzyk J, A measure of similarity for intuitionistic fuzzy sets. In: Proceedings of 3rd conference of the European Society for fuzzy logic and technology, 2003, pp. 206-209.
8. Wen-Liang Hung, Miin-Shen Yang, Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance. Pattern Recognition Letters 25(14) (2004) 1603-1611.
9. Pierpaolo D'Urso, Paolo Giordani A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics & Data Analysis 50(6) (2006) 1496-1523.
10. Wen-Liang Hung, Jinn-Shing Lee, Cheng-Der Fuh, Fuzzy Clustering Based On Intuitionistic Fuzzy Relations. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12(4) (2004) 513-530.
11. Dimitrios K. Iakovidis, Nikos Pelekis, Evangelos E. Kotsifakos, Ioannis Kopanakis, Intuitionistic fuzzy clustering with applications in computer vision. In: Advanced concepts for intelligent vision system. Springer, Berlin, 2008.
12. Ioannis K. Vlachos, George D. Sergiadis, The Role of Entropy in Intuitionistic Fuzzy Contrast Enhancement. Foundations of fuzzy logic and soft computing, Springer, Berlin, 2007.
13. Asuncion A, Newman DJ, UCI Repository of Machine Learning Databases. Irvine, University of California, http://www.ics.uci.eedu/∼mlearn/, 2007.
14. Halkidi M, Batistakis Y, Vazirgiannis M, Cluster validity methods: part I. ACM SIGMOD Record 31(2) (2002) 19-27.