

ENTROPY-BASED INTUITIONISTIC FUZZY C-MEANS CLUSTERING

Truong Quoc Hung*, Nguyen Anh Cuong, Nguyen Dinh Dzung

Abstract: *With the rapid development of the uncertain or hesitant and fuzziness datasets, an entropy-based intuitionistic fuzzy c-means clustering (EIFCM) method is proposed based on the intuitionistic fuzzy sets (IFS) for the clustering problems. Utilizing the advantages of the intuitionistic fuzzy sets and fuzzy sets, which are combined in the proposed method, to overcome some drawbacks of the conventional FCM in handling uncertainties or hesitant and also resolve the fuzziness. Experimental results show that the proposed algorithm is better than the traditional fuzzy clustering algorithms.*

Keywords: Fuzzy sets, Intuitionistic fuzzy sets, Intuitionistic, Fuzzy c-means clustering, Entropy-Based intuitionistic.

1. INTRODUCTION

Clustering technique is applied in many fields such as data mining, pattern recognition, image processing etc. It is used to detect any structures or patterns in the data set, in which objects within the cluster level data show certain similarities. Clustering algorithms have different shapes from simple clustering as k-means and various improvements [1], [2], [3], [4], development of a family of fuzzy c-mean clustering (FCM) [7]. With the framework of fuzzy theory, fuzzy techniques are suitable for the development of new clustering algorithms because they are able to remove vagueness/imprecision in the data [8].

Recently, the intuitionistic fuzzy set (IFS) was introduced [9] and used for representing the hesitance of an expert on determining the membership functions and the non-membership functions. This capability has created a different research direction to handle the uncertainty based on IFS [15]. IFSs also have been recently used for the clustering problem [16]. In [10], the incomplete nutrient-deficient crop images with missing pixels is segmented by an intuitionistic fuzzy clustering algorithm with good results. An other application of the intuitionistic clustering is to evaluate the sport tourism event problem with corporate social responsibility (CSR) and developing a novel Intuitionistic Fuzzy Importance performance Analysis (IFIPA) [11].

Besides, an other type of intuitionistic fuzzy clustering is introduced in [12], the possibilistic intuitionistic fuzzy C-Means clustering algorithm which is the combination of the fuzzy c-means (FCM), possibilistic c-means (PCM) algorithms and intuitionistic fuzzy sets. This algorithm is applied for MRI brain image segmentation with impressive results. In [14], the intuitionistic possibilistic fuzzy c-means clustering algorithm which is proposed to hand the information regarding membership values of objects to each cluster by generalizing membership and non-membership with hesitancy degree.

However, the intuitionistic fuzzy clustering methods which were previously introduced, only based on the various distance and similarity measures among intuitionistic fuzzy sets (IFS) [13] and have difficulties in deciding the most

suitable for measuring the degree of distance or similarity. In addition, they did not care about the entropy of the IFS.

Through the overview of intuitionistic fuzzy clustering presented above, we found an outstanding method of fuzzy sets and intuitionistic fuzzy sets which can be combined in one objective function for handling the hesitant and uncertainties.

Remain of the paper is organized as follows: Section II briefly introduces about some backgrounds about intuitionistic fuzzy sets and fuzzy clustering; Section III proposes the intuitionistic fuzzy C-means clustering algorithm; Section IV offers some experimental results and section V concludes the paper.

2. BACKGROUND

2.1. Intuitionistic Fuzzy sets

Intuitionistic fuzzy sets (IFS) were introduced by Atanassov as an extension of the fuzzy set theory in 1986 as follows:([9]):

Let X be an ordinary finite non-empty set. An IFS in X is an expression \tilde{A} given by: $\tilde{A} = \{x, \mu_{\tilde{A}}(x), \nu_{\tilde{A}}(x) : x \in X\}$ where $\mu_{\tilde{A}} : X \rightarrow [0; 1]$ $\nu_{\tilde{A}} : X \rightarrow [0; 1]$ satisfy the condition $\mu_{\tilde{A}}(x) + \nu_{\tilde{A}}(x) \leq 1$ for all $x \in X$. The numbers $\mu_{\tilde{A}}(x)$ and $\nu_{\tilde{A}}(x)$ denote respectively the degree of membership and the degree of non-membership of the element x in set \tilde{A} . Considering IFSs(x) as the set of all the intuitionistic fuzzy sets in X . For each IFS \tilde{A} in X , The values $\pi_{\tilde{A}}(x) = 1 - \mu_{\tilde{A}}(x) - \nu_{\tilde{A}}(x)$ is called the degree of uncertainty of x to \tilde{A} , or the degree of hesitancy of x to \tilde{A} .

Note that for an IFS \tilde{A} , if $\mu_{\tilde{A}}(x) = 0$, then $\nu_{\tilde{A}}(x) + \pi_{\tilde{A}}(x) = 1$, and if $\mu_{\tilde{A}}(x) = 1$ then $\nu_{\tilde{A}}(x) = 0$ and $\pi_{\tilde{A}}(x) = 0$.

2.2. Entropy on intuitionistic fuzzy sets

Most of the fuzzy algorithms select the best threshold t using the concept of fuzzy entropy. In this paper, we will focus on the definition and characterization of the Intuitionistic fuzzy entropy. The entropy on IFSs is defined as a magnitude that measures the degree of IFS that a set is with respect to the fuzziness of this set which satisfies the following conditions:

1. The entropy will be null when the set is a FSs(x),
2. The entropy will be maximum if the set is an AIFS; that is, $\mu(x) = \nu(x) = 0$ for all $x \in X$,
3. As in fuzzy sets, the entropy of an IFS will be equal to its respective complement
4. If the degree of membership and the degree of non-membership of each element increase, the sum will as well, and therefore, this set becomes fuzzier, and therefore the entropy should decrease. One of the simplest expressions that satisfy the conditions previously mentioned in [17]

$$IE(\tilde{A}) = \frac{1}{n} \sum_{k=1}^n \pi_{\tilde{A}}(x_k) \quad (1)$$

Equation 1 is a base for segmentation algorithm.

2.3. Entropy on intuitionistic fuzzy sets

Fuzzy c-means (FCM) was first introduced by Bezdek in [7]. It is a method of clustering which allows a data point can belong to more than one cluster with

different membership grades. It assumes that a number of clusters c is known in prior and minimizes the objective function (J_m) as:

$$J_m(U, v) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^m (d_{ik})^2 \quad (2)$$

Where:

$$d_{ik} = d(x_k - v_i) = \|x_k - v_i\| = \left[\sum_{j=1}^d (x_{kj} - v_{ij})^2 \right]^{1/2}$$

and m is a constant, known as the fuzzifier, which controls the fuzziness of the resulting partition and can be any real number greater than 1 but generally, it can be taken as 2.

Predefined parameters to the problem: the number of clusters c ($1 < c < n$), fuzzifier m ($1 < m < +\infty$) and error ε . This algorithm can be briefly described as follows:

Algorithm 1: Fuzzy C-means algorithm

1 Step 1: Initialize centroid matrix $V = [v_{ij}]$; $V^{(0)} \in R^{M \times c}$; $j = 0$, by choosing randomly from dataset $X = \{x_i, x_i \in R^M\}$, $i = 1..n$ and the membership matrix U^0 by using the equation:

$$2 \quad u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}}, 1 \leq i \leq c, 1 \leq k \leq n \quad (3)$$

3 Where $d_{ik} = d(x_k - v_i) = \|x_k - v_i\|$ is the Euclidian distance from object x_k to the cluster center v_i

4 Step 2:

5 **Repeat:**

6 Update the centroid matrix $V^{(j)} = [v_1^{(j)}, v_2^{(j)}, \dots, v_c^{(j)}]$ by:

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}, 1 \leq i \leq c$$

7 Update the membership matrix $U^{(j)}$ by using (3)

8 Assign data x_j to cluster c_i if data $(u_i(x_j) > u_k(x_j))$, $k = 1, \dots, c$ and $j \neq k$.

9 **until** : $Max(\|U^{(j+1)} - U^{(j)}\|) \leq \varepsilon$

10 Step 3: Return U and V.

3. ENTROPY-BASED INTUITIONISTIC FUZZY C-MEANS CLUSTERING

As an enhancement of classical FCM, the EntropyBased Intuitionistic Fuzzy C-means Clustering(EIFCM) use the intuitionistic fuzzy sets with the aim to better handle the hesitant data. The general idea of this algorithm is both to minimize the distance from data points to cluster centroids such in FCM and minimize the hesitant to the IFS or entropy on IFS (Eq.1). Next, the hesitant of the IFS is integrated with fuzzy membership in FCM to handle the fuzziness.

The first step is to build the objective function to minimize the hesitant to an intuition of the IFS or entropy on IFS (Eq.1) and the distance between all of the data points to the clusters, simultaneously. Thus, an objective function in IFCM is formed as follows:

$$J_1 = \sum_{i=1}^c \sum_{j=1}^n (1 - \pi_{ij})^p d_{ij}^2 + \sum_{i=1}^c \beta_i \sum_{j=1}^n (\pi_{ij})^p \quad (4)$$

where π_{ij} is the hesitant of the j^{th} data point to the i^{th} cluster, $d_{ij} = ||x_j - v_i||$ is Euclidean distance between the data point x_j and the centroid v_i , c is number of clusters, n is number of data points, p is a weighting exponent for the hesitant membership ($p > 1$), the scale parameter β_i of each cluster.

To minimize the objective function J_1 (Eq.4), the first component with the distance between all of the data points to the cluster centroids is as small as possible, while the second component requires the hesitant of IFS or entropy of IFS is also as small as possible. The scale parameter β_i for each the i^{th} cluster is determined depending on the size of the i^{th} cluster and can be determined as follows:

$$\beta_i = \frac{\sum_{j=1}^n \pi_{ij}^p d_{ij}^2}{\sum_{j=1}^n \pi_{ij}^p} \quad (5)$$

The main idea of the Entropy-Based Intuitionistic Fuzzy C-means Clustering (EIFCM) algorithm is the combination of the fuzzy membership and the hesitant. Thus, the objective function for the hesitant (Eq.4) is considered as a distance measurement and integrated with the fuzzy membership of the FCM, we have the objective function of the EIFCM as follows:

$$J_{EIFCM} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \left((1 - \pi_{ij})^p d_{ij}^2 + \beta_i (\pi_{ij})^p \right) \quad (6)$$

$$= \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \left((1 - \pi_{ij})^p d_{ij}^2 + \sum_{i=1}^c \beta_i \sum_{j=1}^n u_{ij}^m (\pi_{ij})^p \right) \quad (7)$$

in which: u_{ij} is the fuzzy membership of the data point x_j to the i^{th} cluster and m is a constant as in FCM.

Theorem 3.1: The JEIF CM in Eq.7 attains its local minima when $U \equiv [u_{ij}]c \times n$ and $Q = [\pi_{ij}]c \times n$ are assigned the following values:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}^*}{d_{kj}^*} \right)^{1/(m-1)}} \quad (8)$$

$$\pi_{ij} = \frac{1}{1 + \left(\frac{\beta_i}{d_{ij}^2} \right)^{1/(p-1)}} \quad (9)$$

In which $i = 1, 2, \dots, c; j = 1, 2, \dots, n; c$ is the number of clusters and n is the number of patterns and

$$d_{ij}^* = (1 - \pi_{ij})^p d_{ij}^2 + \beta_i \pi_{ij}^p \tag{10}$$

with d_{ij} is the Euclid distance between data point x_j and the centroid v_i which is defined as:

$$v_i = \frac{\sum_{j=1}^n u_{ij}^m (1 - \pi_{ij})^p x_j}{\sum_{j=1}^n u_{ij}^m (1 - \pi_{ij})^p} \tag{11}$$

the scale parameter for each i^{th} cluster is determined depending on the size of the i^{th} cluster can be determined as follows:

$$\beta_i = \frac{\sum_{j=1}^n u_{ij}^m \pi_{ij}^p d_{ij}^2}{\sum_{j=1}^n u_{ij}^m \pi_{ij}^p} \tag{12}$$

The proof of Theorem 3.1 can be easily done with lagrange multiplier with the constraint $\sum_{i=1}^c u_{ij} = 1$

Because of applying the fuzzy membership and the hesitant to the EIFCM algorithm with the constraint $u_{ij} + \pi_{ij} \leq 1$, if $u_{ij} + \pi_{ij} > 1$ then the u_{ij} and π_{ij} will be normalized as

$$t = u_{ij}; u_{ij} = \frac{u_{ij}}{u_{ij} + \pi_{ij}} \text{ and } \pi_{ij} = \frac{\pi_{ij}}{t + \pi_{ij}}$$

In defuzzification step, the combined membership u_{ij}^* is usually as follows:

$$u_{ij}^* = u_{ij} + \pi_{ij} \tag{13}$$

Where: u_{ij} is the membership and π_{ij} is the hesitant of the j^{th} data point to the i^{th} cluster.

However, Eq.13 seems ineffective, we take a simple example as follows: Predefined a data x , a set A . To assess a membership function of x in set A based on intuitionistic fuzzy theory. Let u is the membership function, v is the non-membership function and the hesitation degree π ($u+v+\pi=1$)

Case 1: Assuming that: $u = 0.6, v = 0.1$ and $\pi = 0.3$ we have $u^* = 0.9$ according to Eq.13

Case 2: Assuming that: $u = 0.8, v=0.1$ and $\pi = 0.1$. we also have $u^* = 0.9$ according to Eq.13

One can easily realize that: case 2 is better than case 1. However, with the above Eq.13, the aggregate membership functions u^* are 0.9 in both two cases and we cannot determine the better one.

Remark: The fact that hesitation degree π usually is not affect performance results. We are only interested in two factors that affect the decision are the membership function u and the non-membership function v .

Therefore, we propose a new way to define the aggregate membership functions u^* satisfying properties: the maximum membership function and the minimum

non-membership function

$$u_{ij}^* = u_{ij} - v_{ij} \tag{14}$$

Where: u_{ij} is the membership function Eq.3 and v_{ij} is the the non-membership function of the j th data in i^{th} cluster.

Substituting $v_{ij} = 1 - u_{ij} - \pi_{ij}$ in Eq.14, we have:

$$u_{ij}^* = u_{ij} - v_{ij} \tag{15}$$

$$= u_{ij} - (1 - u_{ij} - \pi_{ij}) \tag{16}$$

$$= 2u_{ij} + \pi_{ij} \tag{17}$$

The experiments are completed for several Machine Learning data (<http://archive.ics.uci.edu/ml/>)

Table 1. Characteristic of heart disease datasets.

Dataset	No of Instance	Class
Cleveland	303	5
Hungarian	294	5
Switzerland	123	5
Long Beach VA	200	5

Table 2. Characteristic of breast cancer diagnostic datasets.

Dataset	No of Instance	Class
Breast-cancer-Wisconsin	699	2
WDBC	569	2
WPBC	198	2

Parameters $m = 2$ for FCM, IFCM, fuzifier $m = 2$; $p = 2$ for EIFCM, the number of clusters c is the number of classes and error parameter $\epsilon = 0:00001$.

The resulting classification performance of the classification is evaluated by the accuracy rate CT as follows:

$$CT = \frac{TR}{TT} \tag{18}$$

Where: TR is the number of correctly classified data and TT is the total number of data points.

The clustering results of the clustering or the quality of classification are reported in terms of index CT, which are shown in Tab.3 and Tab.4

Table 3. Clustering results of heart disease datasets.

Dataset	FCM	IFCM	EIFCM
Cleveland	85.6	87.6	92.1
Hungarian	78.9	85.5	91.4
Switzerland	76.5	90.1	90.1
Long Beach VA	87.8	88.9	94.3

Table 4. Clustering results of breast cancer diagnostic datasets.

Dataset	FCM	IFCM	EIFCM
Breast-cancer-Wisconsin	90.3	87.4	92.8
WDBC	89.8	89.3	93.4
WPBC	77.8	81.7	90.9

These experimental results in Tab.3 and Tab.4 show that the effectiveness of the algorithm EIFCM is better than FCM and IFCM with the accuracy rate CT is always over 90%.

Experiment: The experiments are done based on well-known images with the predefined the number of clusters images in TableV and Fig.1. The results were measured on the basis of several validity indexes to assess the performance of the algorithms on the experimental images.

Table 5. The number of clusters.

Image	Number of cluster
Rose	3
Wolf	3
Mountain	4








Figure 1. Test Images: a) Rose image; b) Wolf image; c) Mountain image.

Table 6. The various validity indexes on the experimental images.

Validity Index	Rose Image			Wolf Image			Mountain Image		
	FCM	IFCM	EIFCM	FCM	IFCM	EIFCM	FCM	IFCM	EIFCM
DB-I	1.4357	1.3252	1.2132	2.3122	1.5631	1.3822	1.5352	1.3312	1.2532
XB-I	0.5281	0.2741	0.1936	0.7214	0.2431	0.1811	0.9854	0.9414	0.8653
S-I	0.9641	0.9172	0.8382	0.9221	0.4943	0.1654	0.7158	0.4261	0.3862
CE-I	0.8432	0.7531	0.6287	0.8131	0.7352	0.6028	0.8643	0.5243	0.5341
PC-I	0.8890	0.8912	0.8912	0.8998	0.9310	0.9413	0.9229	0.9321	0.9591

These test images were clustered by the FCM, IFCM [5] and EIFCM algorithms with predefined parameters: the exponential parameters $m = 2$ for FCM

and IFCM, $m = 2$; $p = 2$ for EIFCM, the number of clusters c is the number of classes and error parameter $\varepsilon = 0.00001$. The different validity indexes is used to assess the clustering results such as the Bezdeks partition coefficient (PC-I), the Dunns separation index (Dunn-I), the Davies-Bouldins index (DB-I), and the Separation index (S-I), Xie and Beni's index (XB-I), Classification Entropy index (CE-I) [6]. The various validity indexes are shown in the Tab.6.

Note that: The validity indexes are proposed to evaluate the quality of clustering. The better algorithm has smaller T-I, DB-I, XB-I, S-I, CE-I and larger PCI. The results in Table 6 show that the EIFCM (the proposed algorithm) have more good performance or higher quality clustering times than the FCM, IFCM algorithms.

4. CONCLUSION

This paper presented a fuzzy c-means clustering algorithm based on intuitionistic fuzzy sets, which improved the clustering results and overcome the drawbacks of the conventional clustering algorithms in handling the hesitant. The proposed approach has solved the problem of combining between IFSs and fuzzy clustering to improve the quality of clustering results. The experiments are done based on some well-known datasets with the statistics show that the proposed algorithm generates better results than the traditional method like FCM. The next goal is some researches related to use the general type-2 intuitionistic fuzzy sets to better improvement of quality.

REFERENCES

- [1]. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, "An Efficient k -Means Clustering Algorithm: Analysis and Implementation", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol 24(7), 881-893, (2002).
- [2]. M.C. Hung, J. Wu, J.H. Chang, D.L. Yang, "An Efficient k -Means Clustering Algorithm Using Simple Partitioning", *J. of Info. Science and Engineering*, Vol.21, 1157-1177, (2005).
- [3]. K.R. Zalik, "An efficient k -means clustering algorithm", *Pattern Recognition Letters* Vol. 29, 1385 - 1391, (2008).
- [4]. K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k -means Clustering Algorithm", *Proceedings of the World Congress on Engineering*, (2009).
- [5]. Z. Xu, J. Wu, "Intuitionistic fuzzy C-means clustering algorithms", *Journal of Systems Engineering and Electronics*, vol.21 (4), pp.580-590, (2010).
- [6]. W. Wang, Y. Zhang, "On fuzzy cluster validity indices", *Fuzzy Sets and Systems* 158, 2095-2117, (2007).
- [7]. J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", *New York: Academic*, (1981).
- [8]. J. Yu and M.-S. Yang, "Optimality test for generalized FCM and its application to parameter selection", *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 1, pp. 164176, (2005).

- [9]. K. Atanassov, "Intuitionistic fuzzy sets", *Fuzzy Sets and Systems*, vol.20, pp. 87-96, (1986).
- [10]. P. Balasubramaniam, V.P. Ananthi, "Segmentation of nutrient deficiency in incomplete crop images using intuitionistic fuzzy C-means clustering algorithm", *Nonlinear Dynamics*, vol.83(1-2), pp. 849-866, (2016).
- [11]. F.H. Huang, Y.J. Ye, C.H. Kao, "Developing a novel Intuitionistic Fuzzy Importance-performance Analysis for evaluating corporate social responsibility in sport tourism event", *Expert Systems With Applications*, vol.42(19), pp.6530-6538, (2015).
- [12]. H. Verma, R.K. Agrawal, "Possibility Intuitionistic Fuzzy c-Means Clustering Algorithm for MRI Brain Image Segmentation", *International Journal on Artificial Intelligence Tools*, vol.24(5), (2015).
- [13]. W.S. Sheng. B.Q. Hu, "Aggregation distance measure and its induced similarity measure between intuitionistic fuzzy sets", *Pattern Recognition Letters*, vol.60-61, pp.65-71, (2015).
- [14]. A. Chaudhuri, "Intuitionistic Fuzzy Possibilistic C Means Clustering Algorithms", *Advances in Fuzzy Systems*, (2015).
- [15]. P.Burillo and H.Bustince, "Construction theorems for intuitionistic fuzzy sets", *Fuzzy Sets and Systems*, vol.84, pp.271-281, (1996).
- [16]. P.Couto, "Image segmentation using atanassov intuitionistic fuzzy sets", Ph.D. thesis, Trs-os-Montes e Alto Douro University, Vila Real, Portugal, (2006).
- [17]. H.Bustince, E.Barrenechea, M. Pagola and R. Orduna, "Image Thresholding Computation Using Atanassovs Intuitionistic Fuzzy Sets", *Journal of Advanced Comput. Int. and Int. Informatics*, vol.11 (2), pp. 187-194, (2007).

TÓM TẮT

PHÂN CỤM C-MEANS MỜ TRỰC CẢM DỰA TRÊN ENTROPY

Với sự phát triển nhanh chóng của các bộ dữ liệu không chắc chắn, một phương pháp phân cụm C-means mờ trực cảm entropy (EIFCM) được đề xuất dựa trên cơ sở các tập mờ trực cảm (IFS) cho các bài toán phân cụm. Các ưu điểm của của các tập mờ trực cảm và các tập mờ được kết hợp với nhau trong phương pháp đề xuất để khắc phục một số hạn chế của phương pháp FCM trong việc xử lý dữ liệu không chắc chắn hoặc do dự cũng như giải quyết tính mờ. Các kết quả thực nghiệm cho thấy thuật toán được đề xuất tốt hơn so với các thuật toán phân cụm mờ truyền thống.

Từ khóa: Tập mờ, Tập mờ trực cảm, Trực cảm, Phân cụm C-means mờ, Trực cảm dựa trên Entropy.

Received 27th February 2018

Revised 5th April 2018

Accepted 20th April 2018

Author affiliations:

Le Quy Don Technical University, Hanoi.

*Corresponding author: truongqhung.@gmail.com