



An equidistance index intuitionistic fuzzy c-means clustering algorithm based on local density and membership degree boundary

Qianxia Ma¹ · Xiaomin Zhu¹ · Xiangkun Zhao¹ · Butian Zhao² · Guanhua Fu³ · Runtong Zhang²

Accepted: 28 January 2024 / Published online: 27 February 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Fuzzy c-means (FCM) algorithm is an unsupervised clustering algorithm that effectively expresses complex real world information by integrating fuzzy parameters. Due to its simplicity and operability, it is widely used in multiple fields such as image segmentation, text categorization, pattern recognition and others. The intuitionistic fuzzy c-means (IFCM) clustering has been proven to exhibit better performance than FCM due to further capturing uncertain information in the dataset. However, the IFCM algorithm has limitations such as the random initialization of cluster centers and the unrestricted influence of all samples on all cluster centers. Therefore, a novel algorithm named equidistance index IFCM (EI-IFCM) is proposed for improving shortcomings of the IFCM. Firstly, the EI-IFCM can commence its learning process from more superior initial clustering centers. The EI-IFCM algorithm organizes the initial cluster centers based on the contribution of local density information from the data samples. Secondly, the membership degree boundary is assigned for the data samples satisfying the equidistance index to avoid the unrestricted influence of all samples on all cluster centers in the clustering process. Finally, the performance of the proposed EI-IFCM is numerically validated using UCI datasets which contain data from healthcare, plant, animal, and geography. The experimental results indicate that the proposed algorithm is competitive and suitable for fields such as plant clustering, medical classification, image differentiation and others. The experimental results also indicate that the proposed algorithm is surpassing in terms of iteration and precision in the mentioned fields by comparison with other efficient clustering algorithms.

Keywords Equidistance index · Local density · Membership degree boundary · Intuitionistic fuzzy c-means · Equidistance index intuitionistic fuzzy c-means

1 Introduction

As an essential branch of machine learning, clustering analysis aims to gather high similarity data samples into the same group. As an unsupervised learning algorithm, clustering has been widely used in many fields, such as image segmentation [1], evaluation of credit risk prediction [2], and pattern recognition [3]. In various clustering algorithms [4–7], the fuzzy c-means clustering (FCM) proposed by Bellman et al.

[8] can integrate the uncertainty of the actual datasets by combining Zada's fuzzy theory [9]. The use of fuzzy information is mostly driven by the ability to understand operations in a manner akin to human logical thinking, which can capture more information about actual problems [10]. In FCM clustering, the interaction between different clusters is generated by FCM, which can effectively avoid falling into the local optimal solution [11, 12]. Due to the uncertainty in data collection in practical problems, FCM may experience uncertainty when calculating the membership value of a given sample [13]. In other words, due to the fact that fuzzy theory only obtains uncertain information through membership functions in expressing fuzzy information, this can result in the loss of some fuzzy information [14]. Therefore, FCM has certain limitations in comprehensively obtaining uncertain information [15].

In order to improve the problem of fuzzy sets being unable to obtain more uncertain information, Atanassov

✉ Xiaomin Zhu
xmzhu@bjtu.edu.cn

¹ School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China

² School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

³ Rail Transit Department, Tianjin Jinhang Computing Technology Research Institute, Tianjin 300308, China

extended Zada's fuzzy theory and proposed the intuitionistic fuzzy set [16]. The IFS uses membership and non-membership functions to describe fuzzy information based on fuzzy sets, to avoid information loss. Scholars also mentioned that IFS allows for correct modeling of the problem based on available data and observation [17]. Given the powerful ability of intuitive fuzziness to capture uncertain information, it has been widely applied in multiple fields, such as fuzzy multi-attribute decision-making problems [18], classification problems [19, 20], forecasting problems [21], and others. Due to the fact that IFSs integrated the concepts of non-membership and hesitation degree in addition to membership in the datasets, which better represents the inherent uncertainty of datasets, IFSs are expanded to FCM [15, 22, 23]. Compared with FCM, IFCM has been proven to converge to better positions and have higher performance in some problems [22, 23]. The IFCM, as a version of FCM, computes the partition matrix or membership matrix by determining the membership value of each data point to join in a cluster, and the cluster centroids are also initialized randomly [15]. Therefore, although IFCM is an improvement for FCM in expressing uncertain information, its initial clustering center still heavily relies on the clustering center [24]. In response to this drawback, the current research mainly focuses on identifying the initial clustering center by obtaining density information on data sample distribution. This idea is first extended to FCM to verify the impact of considering sample distribution density on algorithm performance. Currently, most density-based improved algorithms are implemented by obtaining the initial cluster center by the data points with high density based on cut-off distance [25–28]. To improve the impact of the initial cluster center on the performance of the IFCM algorithm, Varshney et al. [15] also extended the calculation density based on the cut-off distance to IFCM and proposed the density-based IFCM algorithm. The cut-off distance $dc \in [0, 1]$ is defined as a random constant value, which is chosen experimentally. In addition, when calculating the density, the density rate λ , distance rate σ and other parameters need to be adjusted, which makes it difficult to calculate the cut-off density. The above sample density values are unfortunately overdependent on the cut-off distance and heavily affected by noise points.

Apart from the above mentioned shortcomings, the traditional algorithm converges too slowly because all the samples will always affect all the cluster centers [29]. Therefore, scholars have made great efforts to improve the convergence speed of the traditional algorithm. The current improvements can be divided into two types, including improvements to traditional algorithms from the perspective of data features and algorithm principles. From the perspective of data features, the entropy measure [30], distance measure [25, 31], and probabilistic Euclidean distance [26] are extended to

obtain the contributions of different features to the sample. For instance, Cherif et al. [27] proposed the three new interval type-2 fuzzy similarity measures and joined with fuzzy C-means algorithm. An intuitionistic kernelized total Bregman divergence is proposed to measure the difference and the weighted local information is introduced into the objective function [28]. Improvements from the perspective of algorithm principles include using advanced similarity measures instead of Euclidean distance measures when calculating membership, and incorporating heuristic algorithms to avoid falling into local optimization solutions [32, 33]. In addition, some scholars have improved traditional algorithms from two perspectives to accelerate their convergence speed on large datasets. For example, a fuzzy C-means algorithm for optimizing data clustering is proposed by incorporating the typicality function [34]. The random sampling plus extension FCM (rseFCM) obtains the final effective clustering results by taking random samples into the literal FCM (LFCM). Starting from the viewpoint that all samples always affect all clustering centers, Zhou et al. [24] proposed a new membership compression method to achieve fast clustering by scaling membership. For the separation and processing of information, Joaquín Pérez et al. [35] applied the equidistance index (EI) to obtain statistical information about the displacement of centroids at each iteration for reducing the number of calculations needed in the classification step of hard c-means, without significant loss of quality reduction. The algorithm proposed by Joaquín Pérez et al. [35] also verified that the data points can be divided by the EI, and can be achieved better performance. The extension of the EI to FCM is promising research.

According to the above analysis, although IFCM has the ability to express fuzzy information, its clustering performance is unstable because its initial clustering center is randomly selected. It is still a challenge for IFCM to obtain the cluster center in a simple way and ensure the clustering performance. In addition, the convergence of the IFCM algorithm is too slow, because all samples always affect all cluster centers. Therefore, another challenge of IFCM is to reduce the number of iterations while ensuring the clustering performance. In light of this, a novel algorithm named EI-IFCM is proposed to obtain the initial cluster center and reduce the number of iterations effectively from the local density and membership boundary of samples. Both theoretical and empirical studies indicate that EI-IFCM is clear, efficient, and flexible. The main contributions are summarized as follows:

- (1) A new strategy for obtaining initial cluster centers based on local density is proposed. This strategy fully considers the two characteristics of easy calculation and the influence of noise data on the initial clustering center.

- (2) The EI condition is applied to the datasets division, which can provide relevant knowledge about whether the data sample center changes. The EI condition provides theoretical support for the division of data samples. To our knowledge, the study is among the earliest work that applies the equidistance index condition to the IFCM algorithm.
- (3) The boundary value of the membership degree of IFCM is derived. It can increase the membership value of the data samples to the maximum value and terminate the subsequent iteration. The calculation of the membership boundary significantly saves iterations.
- (4) Extensive validation of the EI-IFCM on numerous real-world datasets has been done to demonstrate the superior performance of the algorithm, and the applicability of the algorithm has also been verified.

The rest of this paper is organized as follows. Section 2 presents some basic concepts and new findings of IFS and IFCM. The novel algorithm based on the local density and membership boundary of IFCM is presented in Section 3. In section 4, experiments and sensitivity analysis on some real-world datasets are given. The conclusion of the paper is shown in Section 5.

2 Preliminaries

In this section, the relevant concepts of IFS, IFCM and the equidistance index are reviewed. Based on these basic concepts, the membership boundary of IFCM is derived.

2.1 Intuitionistic fuzzy C means clustering

Atanassov [16] proposed the intuitionistic fuzzy set (IFS) based on membership $\mu_A(x)$ of the fuzzy set by adding the nonmembership $\nu_A(x)$. An IFS A defined on G is given as follows [16].

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

where $\mu_A(x): G \rightarrow [0,1]$, and $\nu_A(x): G \rightarrow [0,1]$ with the condition $0 \leq \mu_A(x) + \nu_A(x) \leq 1$. The hesitation degree of A is expressed as $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \leq 1$.

The intuitionistic fuzzy complement function is first defined by Bustince et al. [36]. Chaira [22] rewrote the intuitionistic fuzzy complement function as $N(x) = (1-x^\alpha)^{1/\alpha}$, where $\alpha > 0$, $N(1) = 0$ and $N(0) = 1$. Chaira [22] also calculated the non-membership of IFS by using the rewritten intuitionistic fuzzy complement function and gave the transformation form of IFS as follows.

$$A = \left\{ x, \mu_A(x), (1 - \mu_A(x)^\alpha)^{1/\alpha} | x \in G \right\} \tag{2}$$

And the hesitation degree is:

$$\pi_{IFS} = 1 - \mu_A(x) - (1 - \mu_A(x)^\alpha)^{1/\alpha} \tag{3}$$

Clustering is described as the process of obtaining different clusters by calculating the membership of data sample points to multiple cluster centers. The IFCM algorithm is an objective function-based clustering algorithm that may optimize the objective function to determine each data sample's membership degree with each clustering center, hence attaining the purpose of automatic clustering. Suppose $X = \{x_1, x_2, \dots, x_n\}$ is the observation sample set and the features of each sample are IFSs in the s -dimension. The objective function of the intuitionistic IFCM algorithm is as follows [23]:

$$J_m(U, V; X) = \sum_{i=1}^c \sum_{j=1}^n \left(\mu_{ij}^{(t)} \right)^m \times \|x_j, v_i^{(t)}\|^2 \tag{4}$$

where c is the number of clusters, m is a fuzzy parameter, $\|x_j, v_i^{(t)}\|^2$ is Euclidean distance measure between v_i (cluster center) and x_j (data points), and μ_{ij} is the membership value of j th data (x_j) in i th cluster.

To minimize the $J_m(U, V; X)$ by using an iterative process based on the following equation.

$$\mu_{ij}^{(t+1)} = \left(\sum_{k=1}^c \left[\frac{\|x_j, v_i^{(t+1)}\|^2}{\|x_j, v_k^{(t+1)}\|^2} \right]^{(1/(m-1))} \right)^{-1} \tag{5}$$

$$\left(\mu_{ij}^{(t+1)} \right)^* = \mu_{ij}^{(t+1)} + \pi_{ij}^{(t+1)} \tag{6}$$

where $\left(\mu_{ij}^{(t+1)} \right)^*$ denotes the membership of the j th data sample in i th cluster under $t+1$ iteration.

Based on the calculation of $\left(\mu_{ij}^{(t+1)} \right)^*$, the clustering center can be updated using the following equation.

$$v_i^* = \frac{\sum_{j=1}^n \left(\mu_{ij}^{(t+1)} \right)^{*m} x_j}{\sum_{j=1}^n \left(\mu_{ij}^{(t+1)} \right)^{*m}} \tag{7}$$

where $i = 1, 2, \dots, c, j = 1, 2, \dots, n$.

The cluster center and membership matrix are updated after each iteration, and the algorithm stops when $\max_{ij} |U_{ij}^{*new} - U_{ij}^{*prev}| < \epsilon$ is satisfied. ϵ is tolerance for the solution accuracy, which has already been set before implementing the clustering task.

2.2 Boundary value and equidistance index

Zhou et al. [29] obtained the membership degree bounds of FCM by rearranging the Euclidean distances from the cluster centers to the data samples in descending order. They designed the boundary value of the membership degree by using the following inequality [29] for the data sample x_j to the nearest cluster center v_i .

$$\frac{1}{c} \leq \mu_{I_j^*}^{(t+1)} \leq \left[1 + (c - 1) \left(\frac{(\|x_j, v_i\|)^{(1)}}{(d_j^{(t+1)})^{(c)}} \right)^{2/(m-1)} \right]^{-1} \tag{8}$$

where $(d_j^{(t+1)})^{(c)}$ denotes the max distance from the j th data point x_j to the i th cluster center $v_i^{(t+1)}$ in $t+1$ iteration, and $I_j^* = \arg \min_{1 \leq i \leq c} \{d_{ij}^{(t+1)}\}$.

According to Eq. (8), the boundary value of the intuitionistic fuzzy membership degree of IFCM can be proposed in Lemma 1.

Lemma 1 For data sample x_j , the boundary membership value of iteration $t+1$ can be calculated by formula (9).

$$0 \leq (\mu_j^{(t+1)})^* \leq 1 - \left(1 - \left[1 + (c - 1) \left(\frac{\|x_j, v_i^{(t+1)}\|^2 / (d_j^{(t+1)})^{(c)}}{(d_j^{(t+1)})^{(c)}} \right)^{2/(m-1)} \right]^{-\alpha} \right)^{(1/\alpha)} \tag{9}$$

Suppose the nearest and the second nearest cluster centers of data object x_j are v_1 and v_2 , respectively. $(d(t+1)j)(1)$ and $(d(t+1)j)(2)$ are the Euclidean distance from x_j to v_1 and v_2 , respectively. The equidistance index (EI) can be expressed as the difference between the two distances. Let v_1 and v_2 be the nearest and the second nearest cluster centers of an object x_j , respectively. Then the EI can be defined as follows [35].

$$\alpha_j^{(t+1)} = abs \left(\|x_j - v_1^{(t+1)}\|^2 - \|x_j - v_2^{(t+1)}\|^2 \right) \tag{10}$$

It is worth noting that $0 \leq \alpha_j^{(t+1)} \leq \|v_1^{(t+1)} - v_2^{(t+1)}\|^2$. It is known that the EI of each data object x_j will change as the iterative process proceeds. There is a high or low likelihood that the cluster of data object x_j will be changed. The equidistance threshold helps to identify the objects with a high likelihood of cluster change. Then the offsets of the clustering center in two iterations, i.e. t and $t+1$ are $m_1 = \|v_1^{(t+1)} - v_1^t\|$ and $m_2 = \|v_2^{(t+1)} - v_2^t\|$, respectively. Then equidistance threshold $\beta^{(t+1)}$ is defined as follows [35].

$$\beta^{(t+1)} = m_1 + m_2 \tag{11}$$

From the above analysis, it can be seen that it is feasible to determine whether the cluster of a data object is going to change in the next iteration by comparing EI with the

equidistance threshold. Therefore, the partition EI for an object x_j can be summarized as follows:

- (1) If $\alpha(t+1)j \leq \beta^{(t+1)}$, the object x_j has a high likelihood of cluster change for the next iteration in clustering;
- (2) If $\alpha(t+1)j > \beta^{(t+1)}$, the object x_j has a low likelihood of cluster change in the next iteration.

Figure 1 can intuitively illustrate the division requirements of partition of EI condition for the object x_m in the process of clustering iteration. As shown in the figure, it is assumed that the objects x_m and x_n are distributed at the positions, and v_1 and v_2 are the nearest and second nearest centers of these two objects, respectively. The parameter changes before and after iteration are shown in the figure. It can be seen from the figure that the center of x_m is not easy to change, while the center of x_n is likely to change.

3 Equidistance Index Intuitionistic Fuzzy C-Means (EI-IFCM) Clustering Algorithm

In this section, we propose a new algorithm named equidistance index intuitionistic fuzzy c-means (EI-IFCM) clustering algorithm by integrating the IFCM and EI. Figure 2 shows the implementation flow chart of the proposed EI-IFCM clustering algorithm. The proposed algorithm mainly includes two parts, namely, the acquisition of the initial cluster center and the update of the membership matrix. In the first part, the local density is fully considered to obtain the initial cluster center. The equidistance index segmentation is applied to the second part to ensure the fast convergence of the proposed algorithm.

3.1 Acquisition of initial clustering centers

The traditional IFCM is sensitive to the selection of the initial clustering center. The cluster centers are often distributed in areas with dense data points, that is, the cluster centers have a large local density within the cluster range [37]. The initial cluster centers should satisfy the following conditions.

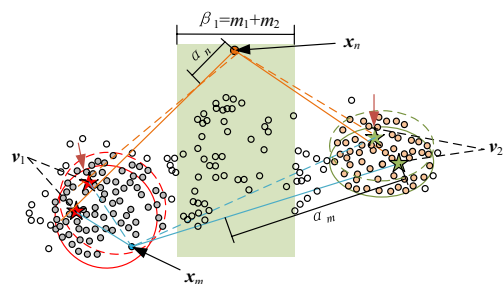
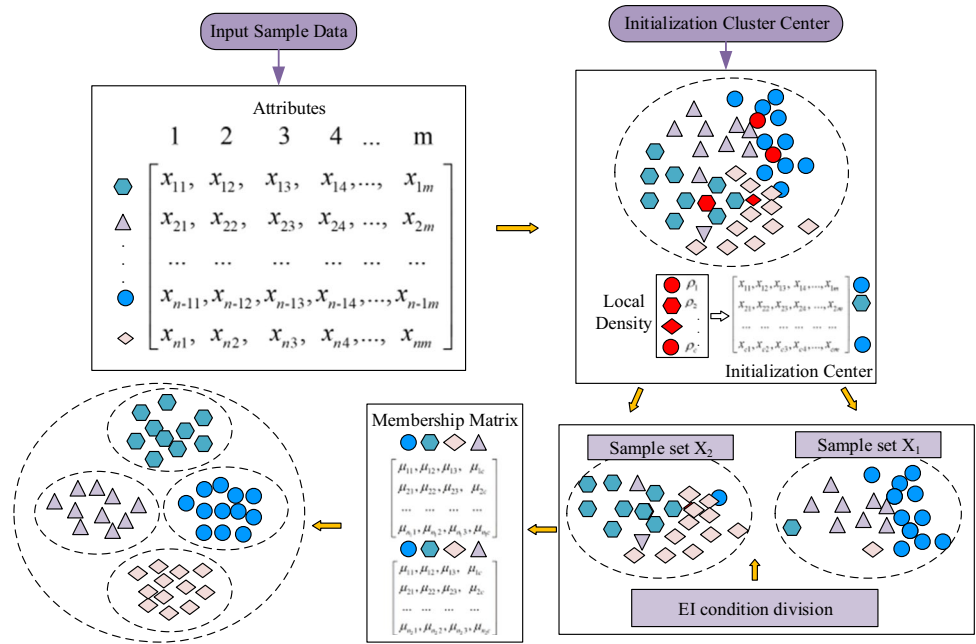


Fig. 1 Schematic diagram of partition of EI condition

Fig. 2 Schematic diagram of the EI-IFCM algorithm implementation process



- (1) If an object v_i is a cluster center, its local density is large;
- (2) If object v_i is the cluster center of a cluster in the data set, the Euclidean distance between the object and the object with a higher local density than it must be large, that is, the object has a larger local density.

Therefore, Algorithm 1 is proposed to select the initial clustering centers in the EI-IFCM algorithm. In Algorithm 1 the global average distance of the dataset is obtained by using E.q (12).

Algorithm 1 Initial center selection

Input: dataset $X = \{x_1, x_2, \dots, x_n\}$.

Output: Initial cluster center $V = \{v_1, v_2, \dots, v_c\}$.

1. Calculate the global average distance \bar{D} of the dataset;
2. Intercept data length k , which satisfied $k = \lambda n$;
3. Calculate Euclidean distance $\sum_{l=1}^k \|x_j, x_l\|^2, j=1,2,3,\dots,n$;
4. Set zero for d_i if $d_g = d_i$; //Avoid duplicate calculations of center values within the same density range.
5. Calculate the local density ρ_j of x_j ;
6. Arrange the local density values in descending order;
7. If the number of cluster centers is less than c , select ρ_j for step 5. And if $\|v_{i-1}, x_j\|^2 \geq \bar{D}$, then $v_i = x_j$, else $j = j + 1$.

$$\bar{D} = \frac{1}{n(n-1)} \sum_{l=1}^k \|x_j, x_l\|^2 \tag{12}$$

And λ is the density factor that can affect the change of local density. The local density can be calculated using E.q (13).

$$\rho_i = e^{\frac{1}{\sum_{l=1}^k \|x_j, x_l\|^2}} \tag{13}$$

where e is an exponential function that is applied to calculate the local density of the data sample point. The exponential

function can eliminate the influence of noise data on the calculation of local density. Therefore, the exponential function is used to calculate the local density of the sample.

Algorithm 1 selects data sample points whose local distance is greater than the global average distance through a descending arrangement of local density. The purpose is to select a large local density and the distance corresponding to two large local density points is greater than or equal to the global average distance to satisfy the initial cluster center selection conditions.

Algorithm 2 EI-IFCM

Input: dataset $X = \{x_1, x_2, \dots, x_n\}$, initial cluster center $\{v_1, v_2, \dots, v_c\}$, fuzzy exponent m , fuzzy complement α and convergence threshold ε ;

Output: Membership matrix U and cluster center matrix V .

1: Call **Algorithm 1** to get the initial cluster center $V^{(0)} = \{v_1, v_2, \dots, v_c\}$.

2: Set $t=1$, update the $U_i^*(t)$ according to equations (3), (5) and (6)

3: Update the $V^{(t)}$ according to **Eq.(7)** and $U_i^*(t)$

4: **Iterate compute stage:**

5: **while** $\varepsilon < 10^{-6}$ and $X_{no} \neq []$ **do**:

$d_{ij}^{(t+1)} = \|x_j, v_i^{(t)}\|^2$ $i=1, 2, \dots, c$ and $j=1, 2, \dots, N$. // Calculate the Euclidean distance $d_{ij}^{(t+1)}$ from data sample X_j to the cluster center $V^{(t)}$.

$(d_{ij}^{(t+1)})^{(1)}, (d_{ij}^{(t+1)})^{(2)}, \dots, (d_{ij}^{(t+1)})^{(c)}, j=1, 2, \dots, N$ // Sort the distance $d_{ij}^{(t+1)}$ in ascending order.

$\alpha_j^{(t+1)} = abs((d_{ij}^{(t+1)})^{(1)} - (d_{ij}^{(t+1)})^{(2)})$ // Calculate EI.

$d_{i(i+1)} = \|v_i^{(t)}, v_{i+1}^{(t)}\|$, $i=1, 2, \dots, c$ // Calculate the distance $d_{i(i+1)}$ between the cluster centers of two adjacent iterations.

Obtain the two distance values $(m_1^{(t+1)}, m_2^{(t+1)})$ with the smallest cluster center shift in two adjacent iterations.

Calculate the equidistance threshold $\beta^{(t)}$: $\beta^{(t+1)} = m_1^{(t+1)} + m_2^{(t+1)}$.

Divide the data sample set by using the equality exponential division condition. */

If $\alpha_j^{(t+1)} > \beta^{(t+1)}$:

$X_{out} = X_j$, $d_{out}^{(t+1)} = d_{ij}^{(t+1)}$, where $j=1, 2, \dots, N(X_{out})$, $i=1, 2, \dots, c$

else

$X_{no} = X_j$, $d_{out}^{(t+1)} = d_{ij}^{(t+1)}$, where $j=1, 2, \dots, N(X_{no})$, $i=1, 2, \dots, c$.

end if

/* Update membership matrix $U^{(t+1)}$ */

If $X_j \in X_{no}$:

$$u_{ij}^{(t+1)} = 1 / \sum_{l=1}^c \left(d_{ij}^{(t+1)} / d_{lj}^{(t+1)} \right)^{2/(m-1)}$$

else if $X_j \in X_{out}$:

/* Sort distance vector $d_{out}^{(t+1)}$, and obtain the minimum and maximum distances of each data sample x_j in X_{out} .

Calculate the membership boundary value **MB** of X_j .

Obtain the minimum distance index I^* from each data sample to the nearest cluster center, where $index = I^*$;

calculate the $U^*(t+1)$ according to Eq.(15). */

If $i = I_j^*$:

$$u_{ij}^{(t+1)} = 1 - \left(1 - \left[1 + (c-1) \left(\|x_j, v_i^{(t+1)}\|^2 / \left(d_j^{(t+1)} \right)^{(c)} \right)^{2/(m-1)} \right]^{-\alpha} \right)^{(1/\alpha)}$$

else

$$u_{ij}^{(t+1)} = u_{ij}^{(t)} \left(\frac{1-M}{1-u_{ij}^{(t)}} \right)$$

end if

end if

Obtain the $U^{(t+1)}$, $V^{(t+1)}$.

Update cluster center vector $V^{(t+1)}$, according to Eq. (7).

Set $X = X_{no}$, $V^{(0)} = V^{(t)}$, $V^{(t)} = V^{(t+1)}$

end while

3.2 Update of membership matrix

The IFCM algorithm updates the membership matrix by calculating the Euclidean distance between the cluster centers and all sample objects. It allows all samples to affect all cluster centers, which will result in some waste of computing resources. When the data samples satisfy the EI division condition, it takes work to shift the cluster in the iteration process. Therefore, the fusion of EI condition and membership boundary values into the IFCM algorithm to update

the membership matrix is the main contribution of the proposed algorithm. The pseudo-code of EI-IFCM based on the EI condition and membership boundary is shown in Algorithm 2. In the proposed algorithm, the data samples satisfying the EI should be separated from the process and assigned the boundary value of the membership.

After obtaining the initial cluster center in Algorithm 1, Algorithm 2 starts its learning process from a better cluster center. In the process of updating the membership matrix, the dataset is divided according to the EI division condition. The

division rules are shown in Eq. (15). The boundary value of membership can be obtained based on Eq. (14) as follows.

$$M = 1 - \left(1 - \left[1 + (c - 1) \left(\frac{\left(\|x_j, v_i^{(t+1)}\| \right)^{(1)}}{\left(d_j^{(t+1)} \right)^{(c)}} \right)^{2/(m-1)} \right]^{-\alpha} \right)^{1/\alpha} \tag{14}$$

Updating the membership of the separated sample X_{out} data by the following formula.

$$u_{ij}^{(t+1)} = \begin{cases} M, & \text{if } i = I_j^* \\ u_{ij}^{(t)} \left(\frac{1-M}{1-u_{ij}^{(t)}} \right), & \text{if } i \neq I_j^* \end{cases} \tag{15}$$

where $\frac{1-M}{1-u_{ij}^{(t)}}$ is a factor less than 1 to maintain the $u_{ij}^{(t)}$ is decreased to $u_{ij}^{(t+1)}$ for $i \neq I_j^*$.

4 Experimental results

4.1 Environment settings and measuring indexes

Nine datasets from the UCI datasets (<https://archivr.ics.uci.edu/ml/index.php>) are utilized in this experiment to verify the competitiveness and effectiveness of the proposed algorithm. The nine datasets mentioned above are aggregated from different groups in fields such as animals, plants, aviation, healthcare etc. In this section, the applicability of the method is demonstrated by analyzing these actual datasets. A personal computer running Windows 10 and Python 3.8.8 is used for all of the experiments. It has an Intel Core i5-1135G7 processor and a maximum memory capacity of 16 GB for all processes. The brand new algorithm is benchmarked with its counterparts such as FCM [8], LFCM [38], IFCM [23], rseFCM [38], MSFCM [29], FCM clustering by varying the fuzziness parameter (vFCM) [39], Improved FCM (IMFCM) [33], and feature weighted FCM (WFCM) [25]. As a classical clustering algorithm, the traditional FCM algorithm is often used to compare the improved clustering algorithms based on FCM. The proposed clustering algorithm integrates the advantages of the IFCM algorithm, and IFCM is also used to compare with our algorithm. The MSFCM applies the boundary value of membership degree to the clustering method for the first time and is used to compare with EI-IFCM. In addition, other methods such as LFCM, rseFCM, IMFCM, WFCM, and vFCM are also used to compare with our methods because of their wide application in clustering in very big datasets. The basic parameters m and α are set to 2 and 0.85, respectively. In addition, the iteration termination parameter ϵ is set as 1×10^{-6} .

The Rand index (RI) for evaluating clustering methods quality proposed by Rand in 1971 is widely used to measure the similarity of two clustering partitions [40]. Since RI was proposed, it has been widely used in the performance evaluation of clustering algorithms. The larger the RI, the better the clustering performance. And the RI is extended by Hubert et al., which is adjusted rand index (ARI) [41]. The ARI inherits the advantages of RI and its value is between [-1, 1]. The Adjusted Mutual Information (AMI) is an information theoretic measure and has the best properties among all these clustering evaluation measures [42]. In addition, the FMI considers the Fowlkes and Mallows Index (FMI) proposed by Fowlkes and Mallows in 1983 [43]. FMI is described as the geometric mean of accuracy and recall, which is used to comprehensively evaluate clustering performance. The larger the value of the above evaluation indexes, the closer the clustering result is to the true value, that is, the better the clustering performance.

4.2 Results and discussion

Table 1 demonstrates the results of FCM, LFCM, IFCM, rseFCM, MSFCM, WFCM, IMFCM, vFCM, and EI-IFCM over nine datasets coming from UCI datasets. The characteristic dimension (d), data amount (n) and category number (c) of the data set are listed in Table 1. On nine datasets, nine algorithms were tried ten times, and the average and standard deviation of each indicator’s ten tests are taken. The standard deviation is roughly 0 when it is less than or equal to 10^{-7} . Specifically, when using the WFCM algorithm, its parameter $t=-1$, and the parameters a, b , and K of the vFCM algorithm are 0.95, 0.05, and 20, respectively. The selection of parameters for the above two algorithms comes from the optimal parameters validated in [25, 39].

In this section, we use the above evaluation indicators mentioned in section 4.1 to measure the clustering performance of algorithms. Table 1 lists the detailed content of each indicator, with the best result highlighted in bold. In Table 1, in addition to evaluating the performance of a single indicator on different datasets, a comprehensive evaluation is also provided. The “Mean” in the table displays the comprehensive performance of the six indicators on all datasets, as well as the average ranking under all indicators in all datasets. Through verification, it can be found that the proposed algorithm can be well used in distinguishing different categories of plants, classifying diseases in medical treatment, and recognizing objects in aerial images and others.

From Table 1, it can be seen that different algorithms exhibit different performances on different datasets. For traditional algorithms such as FCM and LFCM, the convergence position of the algorithm is learned by randomly initializing the membership matrix and randomly initializing the clustering centers. Compared to other algorithms, they

Table 1 The calculation results of 9 algorithms on 9 real datasets

Datasets	Details of datasets	Evaluate criteria	FCM	LFCM	IFCM	rsFCM	MSFCM	WFCM	IMFCM	vFCM	EHFCM		
Iris	$n=150$ $d=4$ $c=3$	ARI	0.7294±0	0.7294±0	0.7430±0	0.7294±0	0.7424±0	0.6422±0	0.6893±0.134	0.7942±0	0.8019±0		
		AMI	0.7465±0	0.7465±0	0.7547±0	0.7465±0	0.7487±0	0.7465±0	0.6733±0	0.7203±0.068	0.7465±0	0.7934±0	
		FMI	0.8197±0	0.8197±0	0.8285±0	0.8197±0	0.8277±0	0.8197±0	0.8277±0	0.8009±0.071	0.8196±0	0.8673±0	
		RI	0.8797±0	0.8797±0	0.8860±0	0.8797±0	0.8859±0	0.8797±0	0.8415±0	0.8569±0.071	0.8797±0	0.9124±0	
		Time/s	0.229±0.021	0.186±0.075	0.004±0.001	0.018±0.002	0.023±0.003	0.018±0.002	0.023±0.003	0.167±0.020	0.006±0.001	0.009±0.002	0.003±0.001
		Iterations	42.4±2.914	34.2±12.80	34±1.63	28.2±4.87	14.7±1.55	28.2±4.87	14.7±1.55	30.5±3.38	50.4±8.64	80±0	4.7±0.82
		ARI	0.3539±0	0.3539±0	0.3615±0	0.3539±0	0.3690±0	0.3539±0	0.3690±0	0.3833±0	0.3907±0	0.3711±0	0.4031±0
		AMI	0.4106±0	0.4106±0	0.4149±0	0.4106±0	0.4203±0	0.4106±0	0.4203±0	0.4291±0	0.4280±0	0.4227±0	0.4331±0
		FMI	0.5728±0	0.5728±0	0.5780±0	0.5728±0	0.5830±0	0.5728±0	0.5830±0	0.5922±0	0.5954±0	0.5836±0	0.6073±0
		RI	0.7105±0	0.7105±0	0.7238±0	0.7105±0	0.7171±0	0.7105±0	0.7171±0	0.7237±0	0.7281±0	0.7186±0	0.7313±0
Wine	$n=178$ $d=13$ $c=3$	Time/s	0.436±0.028	0.454±0.057	0.271±0.025	0.024±0.003	0.016±0.004	1.044±0.068	0.011±0.002	0.069±0.007	0.003±0.001		
		Iterations	67.4±4.52	59±5.67	57.7±5.27	47.2±13.08	18.8±2.78	47.2±13.08	62.4±2.25	71.3±7.33	560±0	5±1.10	
		ARI	0.4914±0	0.4914±0	0.4914±0	0.4914±0	0.5124±0	0.4914±0	0.4883±0.005	0.5834±0	0.4811±0	0.5779±0	
		AMI	0.4640±0	0.4640±0	0.4640±0	0.4640±0	0.4805±0	0.4640±0	0.4616±0.004	0.5204±0	0.4559±0	0.5158±0	
		FMI	0.7915±0	0.7915±0	0.7915±0	0.7915±0	0.7984±0	0.7915±0	0.7905±0.002	0.8215±0	0.7882±0	0.8195±0	
		RI	0.7504±0	0.7504±0	0.7504±0	0.7504±0	0.7605±0	0.7504±0	0.7489±0.002	0.7946±0	0.7454±0	0.7919±0	
		Time/s	0.415±0.026	0.422±0.044	0.300±0.016	0.071±0.006	0.171±0.011	0.071±0.006	2.265±0.104	0.034±0.005	0.053±0.012	0.009±0.002	
		Iterations	33.2±0.79	30.1±2.13	29.1±0.99	29.1±1.85	17.1±0.88	29.1±1.85	29±1.25	172.9±16.27	200±0	2.6±0.84	
		ARI	0.5094±0	0.5094±0	0.5091±0	0.5094±0	0.4994±0	0.5094±0	0.5305±0	0.4264±0.0338	-	0.5254±0	
		AMI	0.5832±0	0.5832±0	0.5675±0	0.5832±0	0.5320±0	0.5832±0	0.6049±0	0.5339±0.025	-	0.5759±0	
Satim-age	$n=6435$ $d=36$ $c=6$	FMI	0.6001±0	0.6001±0	0.6005±0	0.6001±0	0.5957±0	0.6173±0	0.5312±0.028	-	0.6167±0		
		RI	0.8522±0	0.8522±0	0.8514±0	0.8522±0	0.8445±0	0.8522±0	0.8588±0	0.8290±0.010	-	0.8526±0	
		Time/s	80.166±6.267	66.150±8.090	64.575±8.340	34.771±0.128	13.734±2.886	34.771±0.128	396.137±29.358	1.644±0.635	-	4.956±0.131	
		Iterations	116.5±5.07	112.3±19.49	196±23.12	109±5.38	80±38.60	109±5.38	130.1±7.87	378.3±144.34	-	14±1.33	
		ARI	0.0305±0	0.0305±0	0.0363±0	0.0307±0	0.0458±0	0.0307±0	0.0372±0.038	0.0426±0.043	0.0303±0	0.0458±0	
		AMI	0.0187±0	0.0187±0	0.0230±0	0.0212±0	0.0302±0	0.0212±0	0.0235±0.0275	0.0318±0	0.0174±0	0.0302±0	
		FMI	0.5224±0	0.5223±0	0.5241±0	0.5172±0	0.5274±0	0.5172±0	0.5277±0.021	0.5225±0	0.5291±0	0.5274±0	
		RI	0.5154±0	0.5154±0	0.5182±0	0.5154±0	0.5229±0	0.5154±0	0.5187±0.019	0.5213±0	0.5154±0	0.5229±0	
		Time/s	0.240±0.014	0.211±0.023	0.164±0.010	0.081±0.003	0.057±0.007	0.081±0.003	1.230±0.060	0.013±0.001	0.098±0	0.002±0.001	
		Iterations	34.9±1.29	31.3±1.89	32.6±1.78	32.5±5.30	18.6±2.12	32.5±5.30	75.8±3.39	90.1±5.55	920±0	2.9±0.74	
Seeds	$n=210$ $d=7$ $c=3$	ARI	0.7166±0	0.7166±0	0.7166±0	0.7166±0	0.7324±0	0.7031±0	0.7056±0	0.7166±0	0.7647±0		
		AMI	0.6922±0	0.6922±0	0.6922±0	0.6852±0	0.7180±0	0.6852±0	0.6647±0	0.6764±0	0.6922±0	0.7297±0	
		FMI	0.8106±0	0.8106±0	0.8106±0	0.8105±0	0.8215±0	0.8105±0	0.8012±0	0.8032±0	0.8106±0	0.8426±0	
		RI	0.8744±0	0.8744±0	0.8744±0	0.8744±0	0.8812±0	0.8744±0	0.8686±0	0.8695±0	0.8744±0	0.8957±0	
		Time/s	0.283±0.025	0.227±0.029	0.005±0.002	0.019±0.005	0.004±0.002	0.019±0.005	0.339±0.023	0.010±0.0106	0.004±0.001	0.003±0.001	
		Iterations	39.5±2.59	30.2±1.99	39±1.70	30±2.91	15.5±1.72	30±2.91	32.7±1.77	41.3±5.36	34.9±1.20	5.5±1.27	

Table 1 (continued)

Datasets	Details of datasets	Evaluate criteria	FCM	LFCM	IFCM	rsIFCM	MSFCM	WFCM	IMFCM	vFCM	EI-IFCM	
Abalone	$n=4177$ $d=8$ $c=3$	ARI	0.1331±0	0.1331±0	0.1336±0	0.1335±0	0.1543±0	0.1331±0	0.1336±0	0.1331±0	0.1390±0	
		AMI	0.1289±0	0.1289±0	0.1296±0	0.1289±0	0.1399±0	0.1289±0	0.1251±0	0.1289±0	0.1298±0	
		FMI	0.4611±0	0.4611±0	0.4600±0	0.4611±0	0.4546±0	0.4611±0	0.4611±0	0.4231±0	0.4611±0	0.4335±0
		RI	0.5880±0	0.5880±0	0.5893±0	0.5880±0	0.6118±0	0.5880±0	0.6118±0	0.6144±0	0.5880±0	0.6124±0
		Time/s	18.283±1.083	11.634±0.762	0.074±0.006	5.985±0.015	2.286±0.362	17.041±0.555	2.286±0.362	17.041±0.555	0.163±0.015	0.178±0.008
Avila	$n=20867$ $d=10$ $c=12$	Iterations	96.2±3.19	60.2±4.76	77.7±2.11	63.5±3.60	21.6±3.60	62.4±1.90	297.7±22.41	179±0	6±1.70	
		ARI	0.0054±5.31e-6	0.0054±0.004	0.0255±2.30e-5	0.0122±0.003	0.0260±0.012	0.0138±0	0.0260±0.012	0.0118±0.002	-	0.0305±0
		AMI	0.0382±3.86e-6	0.0381±0.006	0.0374±1.11e-5	0.0420±0.006	0.0399±0.003	0.0551±0	0.0399±0.003	0.0498±0.002	-	0.0428±0
		FMI	0.2655±4.83e-7	0.2655±0.024	0.3582±3.76e-5	0.2198±0.022	0.3301±0.024	0.1548±0	0.3301±0.024	0.1890±0.008	-	0.2579±0
		RI	0.6116±8.21e-6	0.6116±0.027	0.5125±5.07e-5	0.6182±0.024	0.5545±0.023	0.7269±0	0.5545±0.023	0.6983±0.005	-	0.6518±0
Shuttle	$n=58000$ $d=9$ $c=7$	Time/s	173.976±13.599	235.259±206.070	315.166±37.669	11431.032±38.648	25.878±0.878	12893.314±1819	366.454±213.680	19880.2±13575.30	-	24.2±2.40
		Iterations	15452±690.13	14711±627.36	128.1±17.33	18358.1±403.65	101.00±19	1779.2±173.84	101.00±19	0.1264±0.028	-	0.1955±0
		ARI	0.1138±0.001	0.1185±0.015	0.1187±2.21e-4	0.1406±0.025	0.1715±0.018	0.1635±0	0.1715±0.018	0.2602±0.032	-	0.2855±0
		AMI	0.2433±0.001	0.2461±0.009	0.2507±7.90e-5	0.2576±0.03	0.2303±0.042	0.2601±0	0.2303±0.042	0.4591±0.029	-	0.5371±0
		FMI	0.4457±0.001	0.4512±0.018	0.4517±2.91e-4	0.4804±0.03	0.5466±0.023	0.4978±0	0.5466±0.023	0.4863±0.019	-	0.5378±0
Mean	All datasets	Time/s	1693.814±455.310	1372.696±349.001	2232.921±1026.159	1019.845±23.129	1067.133±385.565	2157.615±162.192	10.529±2.790	316.1±123.06	-	19±2.16
		Iterations	298.3±47.98	227±54.80	926.2±103.32	187.5±53.84	146±33.85	222.5±8.66	146±33.85	0.3455	-	0.3871
		ARI	0.3426	0.3431	0.3484	0.3464	0.3615	0.3484	0.3615	0.3718	-	0.3929
		AMI	0.3695	0.3698	0.3704	0.3710	0.3711	0.3703	0.3711	0.5718	-	0.6121
		FMI	0.5877	0.5883	0.6003	0.5859	0.6094	0.5132	0.6094	0.7109	-	0.7232
Average Rank	Win/All	RI	0.6955	0.6959	0.6875	0.6986	0.7017	0.7131	0.7109	-	-	9.322
		Time/s	218.649	187.471	290.387	1387.983	123.256	1718.795	123.256	42.096	-	12.958
		Iterations	1797.822	1699.478	168.933	2098.344	48.144	269.400	48.144	2366.478	-	1.648
		Average Rank	5.537	5.111	4.407	4.444	3.130	4.500	4.389	5.807	1.648	
		Win/All	1/54	1/54	2/54	1/54	6/54	7/54	8/54	2/54	32/54	

Bold indicates best results. The basic parameter settings ϵ , m and α are set 1×10^{-6} , 2 and 0.85, respectively

are not competitive on any dataset. The rseFCM algorithm, which integrates LFCM, has been proposed for clustering problems on large datasets. However, it inherits the inherent drawbacks of LFCM, such as randomization of initial centers and slow convergence speed, and its accuracy has not been significantly improved on all datasets. The IFCM algorithm is based on an improvement of the FCM algorithm in considering more uncertain information, which adds hesitation to the membership function of the FCM. Although the IFCM algorithm is random in obtaining the initial clustering centers, it performs well on the Shuttle, Avila, Abalone, Heart, Wine, and Iris datasets due to its consideration of more uncertain information in the data samples.

Compared with the algorithms mentioned above, the MSFCM algorithm performs better on multiple datasets. This is because the MSFCM algorithm proposes the concept of membership boundary from the perspective that all data objects always affect all clustering centers, and integrates it with the membership update process to obtain better and faster convergence positions. From Table 1, it can be seen that MSFCM exhibits good performance on most datasets, which proves that considering the boundary condition of membership is effective for improving the traditional algorithm FCM. The WFCM algorithm is an algorithm that considers feature weighting and the calculation of feature weights depends on the Euclidean distance between the cluster center and the mean of all cluster centers. It can be seen that the datasets Satimage and Avila are sensitive to feature weighting and exhibit good performance, but the performance of the WFCM algorithm on other datasets is mediocre. The IMFCM algorithm takes into account the disadvantage of using Euclidean distance to calculate the membership matrix of the FCM algorithm, which is sensitive to noisy data. It proposes the Euclidean distance function that ensures a significant weight on normal data while a small weight on noisy data. It can be seen that this improvement has shown good performance on the Cancer dataset.

The vFCM algorithm improves FCM from the perspective of selecting and updating fuzzy parameters. VFCM determines whether to use the function $m=am_0+b$ to update fuzzy parameters by determining the remainder of the iteration times t and K . It can be known that when t is a multiple of K , the value of m will decrease. For datasets that require multiple iterations to reach the convergence position, the value of m will decrease to 1 (with an initial value of 2). When the value of the fuzzy parameter is 1, it will cause confusion in the updating of membership due to the limitation of $1/(m-1)$. Therefore, the vFCM algorithm does not converge on some datasets, such as Satimage, Avila, and Shuttle.

The proposed EI-IFCM algorithm integrates the further consideration of uncertain information in membership calculation in IFCM, as well as the use of membership boundary

values in algorithm iterations. In addition, the proposed algorithm also considers the selection of initial clustering centers, which is a gap in the aforementioned algorithms. From Table 1, it can be seen that the EI-IFCM algorithm, which integrates multiple advantages, has shown good performance on most datasets. The “mean” row of the table also fully reflects the overall performance of the proposed algorithm on all datasets, which is competitive.

Additionally, it can be seen that our algorithm consumes less time and iterations. This is because the proposed algorithm first starts its learning from better initial clustering centers, which are obtained based on the density of data samples. The proposed algorithm divides the data samples during the learning process based on the EI condition and assigns boundary values for samples that satisfy the EI condition. By starting the learning process from a better location and assigning membership boundary values that match EI condition, the convergence position is achieved quickly during the algorithm operation process. From Table 1, it can be seen that this viewpoint has been validated on almost all datasets.

To verify whether there are statistical differences between algorithms, we include the p -values obtained by the Wilcoxon test for every indicator of the proposed EI-IFCM algorithm and the other algorithms in Table 2. In addition, Table 2 also summarizes whether different indicators under each algorithm are greater than 0.05 (-) or less than 0.05 (+), and counts their quantities. Specifically, due to the non-convergence of the vFCM algorithm on the Satimage, Shuttle, and Avila datasets, these datasets have been removed when comparing EI-IFCM and vFCM algorithms. Although the proposed algorithm has significant differences in five or four indicators for FCM, IFCM, LFCM, and MSFCM, the p -value is still greater than 0.05 on one or two indicators, indicating that the null hypothesis cannot be rejected. Similarly, compared to the recently proposed new algorithms IMFCM, WFCM, and vFCM, overall, the null hypothesis cannot be rejected. Therefore, we can conclude that there is no statistically significant difference between our algorithm and the most recent algorithms.

Figure 3 shows the visualization results of original samples, IFCM and EI-IFCM algorithms on Wine and Iris data sets. (a) and (d) are 2D figures on original samples. (b) and (c) are the clustering results of the two algorithms on the Wine dataset, respectively. (a) and (c) are the clustering results of the two algorithms on the Iris dataset respectively. It is clear from Fig. 3 that the algorithm proposed in this paper can achieve a better clustering effect than IFCM on Wine and Iris datasets. This shows that the proposed algorithm is competitive. Especially on the Wine dataset, there are more obvious clustering results.

To verify the monotonicity of algorithms, Fig. 4 shows the iterative changes of nine algorithms on nine data sets

Table 2 Wilcoxon test *p*-values obtained by the Wilcoxon test of all evaluate criteria pairs of the proposed algorithm and other algorithms

Evaluate criteria	FCM	LFCM	IFCM	rseFCM	MSFCM	IMFCM	WFCM	vFCM
ARI	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0251(+)	0.0195(+)	0.0078(+)	0.0313(+)
AMI	0.0195(+)	0.0195(+)	0.0195(+)	0.0195(+)	0.0251(+)	0.3594(-)	0.1641(-)	0.0313(+)
FMI	0.0547(-)	0.0547(-)	0.2500(-)	0.0195(+)	0.4817(-)	0.0078(+)	0.0742(-)	0.2188(-)
RI	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0117(+)	0.2500(-)	0.2031(-)	0.0313(+)
Time	0.0039(+)	0.0039(+)	0.0273(+)	0.0039(+)	0.0977(-)	0.9102(-)	0.0039(+)	0.4375(-)
Iterations	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0039(+)	0.0313(+)
+/-	5/1	5/1	5/1	6/0	4/2	3/3	3/3	4/2

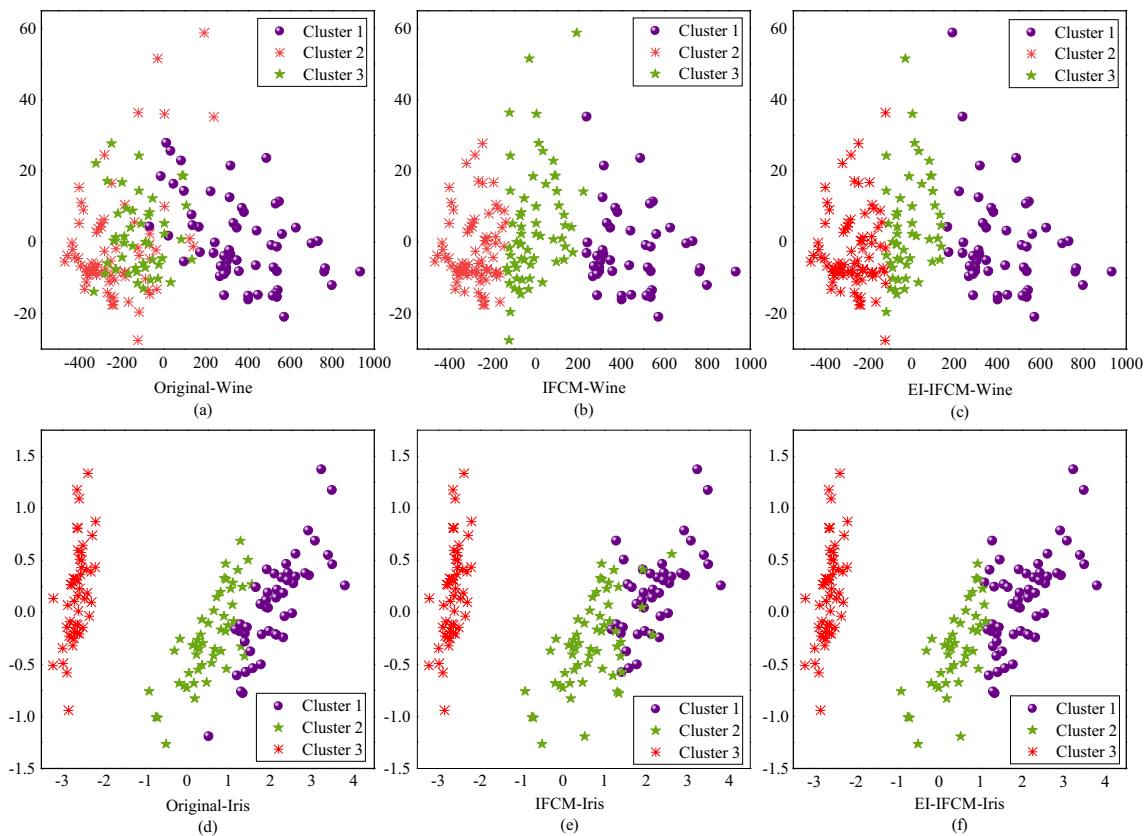


Fig. 3 Plots of Iris and Wine Datasets in 2-D by IFCM and EI-IFCM Algorithm

during 30 iterations. The vertical coordinate represents the sum of the objective function and the initial objective function, i.e. $J(U, V)/J(U_0, V_0)$, and the horizontal coordinate represents the number of iterations. Due to significant differences in target values among different algorithms, using this ratio of $J(U, V)$ and $J(U_0, V_0)$ can plot the results of all algorithms on the same graph. The initial objective values of FCM, LFCM, IFCM, rseFCM, MSFCM, FWFCM, IMFCM and vFCM algorithms are determined by randomly selected cluster centers or randomly given

membership values. The change in the ratio during the iteration process will be significant when the original objective value is large, otherwise, the change will be small. The initial objective value of the EI-IFCM algorithm is determined based on the selection of cluster centers. In addition, it can be seen from Fig. 4 that all datasets changed dramatically at the beginning of the iteration, then gradually slowed down. Compared with other classic algorithms, the EI-IFCM algorithm can achieve convergence with fewer iterations.

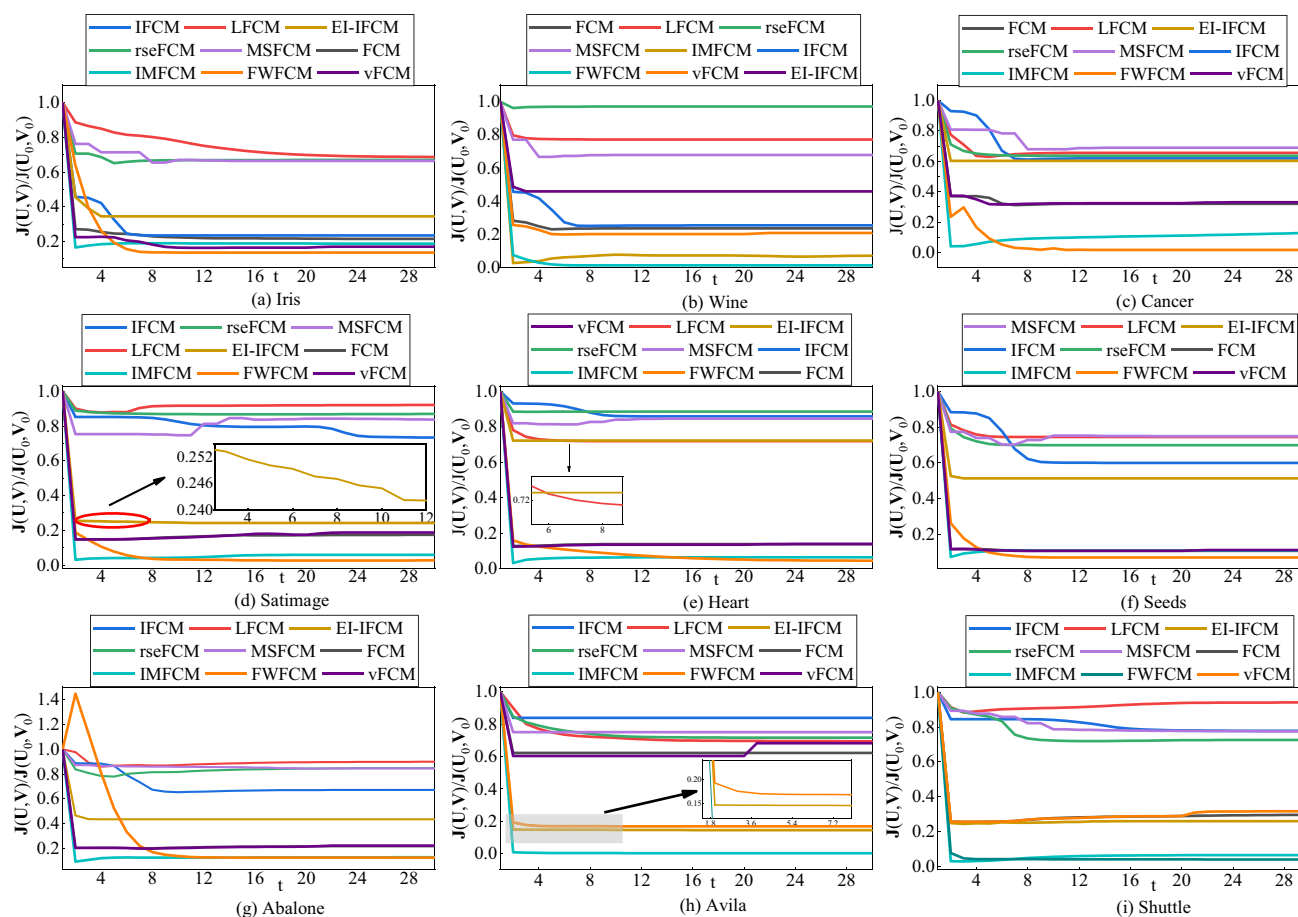


Fig. 4 Plots of $J(U,V)/J(U_0,V_0)$ for the iterations of nine algorithms on nine datasets

4.3 Sensitivity analysis

The stability of the initial clustering center is depicted in Fig. 5. Two representative datasets, Seeds and Satimage, are selected to describe this change. This is because there are significant differences in sample size, feature dimension, and clustering number between the two datasets. The abscissa in Fig. 5 is the number of experiments, and the ordinate is the impact of the initial center on the accuracy of the algorithms. The initial cluster center of IFCM is randomly selected. It can be seen from the figure that the initial centers randomly selected will cause the RI of the algorithm to fluctuate greatly, while the initial clustering centers selected through local density contribute to the stability of the algorithm and ensure the clustering accuracy.

Similar to traditional FCM and IFCM algorithms, the parameters of the EI-IFCM algorithm have an impact on clustering performance. In this section, a detailed analysis is conducted on the impact of each parameter.

To analyze the sensitivity of the proposed algorithm to the fuzzy parameter m , the RI values of FCM and IFCM algorithms under different parameters m in four datasets are

also measured. Figure 6 shows the changes in EI-IFCM, FCM and IFCM algorithms on m . It can be seen from the figure that the RI of the three algorithms fluctuates on different m . This shows that different datasets have different sensitivity to the fuzzy parameter m , and it is necessary to select appropriate fuzzy parameters according to different datasets to obtain higher clustering accuracy. On the datasets of Seeds, Avila and Abalone, the derived algorithm has roughly the same fluctuation as FCM and IFCM. When m is greater than 2.4, IFCM and EI-IFCM show a large gap with FCM in volatility but EI-IFCM and IFCM have roughly the same volatility trend on the Satimage dataset. This is because IFCM and EI-IFCM are affected by another hesitation parameter α besides parameter m .

EI-IFCM and IFCM algorithms are affected not only by fuzzy parameters m but also by the parameter α . Parameter α can change the hesitation value of data samples. Figure 7 shows the impact of IFCM and EI-IFCM algorithms on the RI of datasets when parameters α in $[0.15, 0.95]$ are considered. It can be seen from the figure that IFCM and EI-IFCM fluctuate with the change of parameter α . The fluctuation trend of RI obtained by two algorithms in

Fig. 5 Influence of initial centers selection of IFCM and EI-IFCM Algorithms. (a) Seeds dataset, (b) Satimage dataset

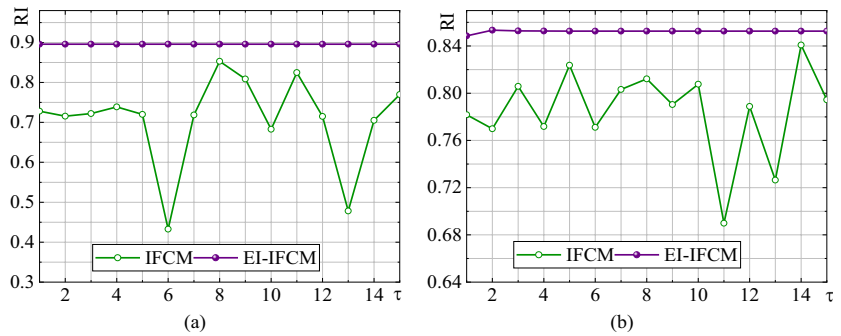
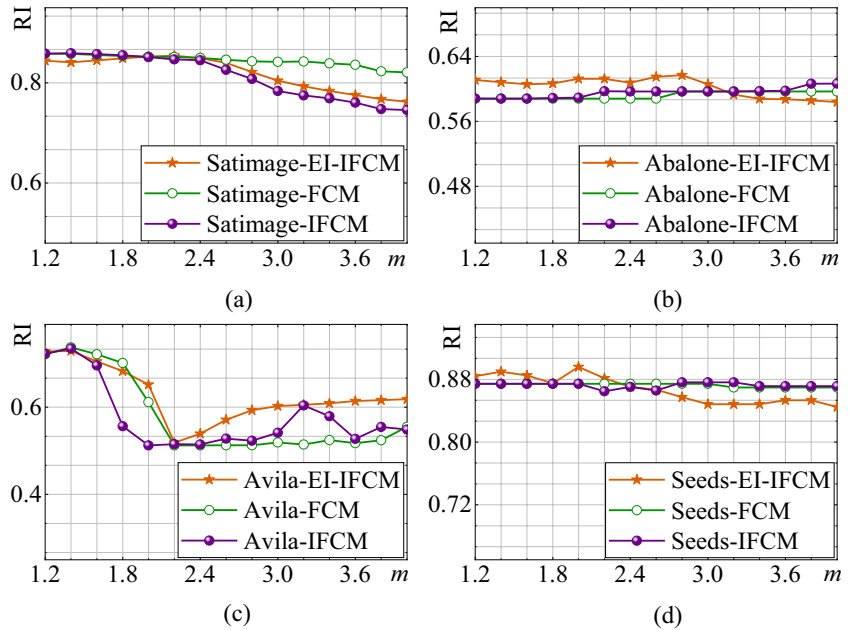


Fig. 6 Plot of hesitation parameter m on clustering performance of three algorithms



different data sets is basically synchronous and consistent. It can be seen from the figure that when $\alpha=0.85$, the two algorithms achieve the highest clustering accuracy in the Seeds and Satimage datasets at the same time. However, when the two algorithms achieve the highest clustering accuracy on the Avila and Abalone datasets, they are not under the same parameters. The IFCM achieves the highest clustering precision on two Avila when $\alpha=0.25$, and EI-IFCM achieves the highest clustering precision on it when α is 0.7. The Abalone dataset achieves the highest clustering precision when $\alpha=0.65$ in IFCM while the highest clustering precision is achieved in EI-IFCM when $\alpha=0.8$. It is worth noting that the RI is 0.6131 in EI-IFCM and 0.6121 in IFCM when $\alpha=0.65$, while RI in EI-IFCM is significantly greater than in IFCM when $\alpha=0.8$ in the Abalone dataset. The same results can be obtained in the Avila dataset. On the whole, the performance of the EI-IFCM algorithm is better than that of the IFCM algorithm, which is most obvious on the Avila dataset, and the excellent ability of the algorithm has also been further proved.

In order to verify the impact of both parameters on algorithm performance, Figure 8 depicts the RI values obtained after both parameters are changed simultaneously. Among them, the value range is α in $[0.1,1]$, and the value range of m is in $[1.2,3.9]$. The x-axis in the figure is α and m , respectively, and the y-axis is the RI value. From the figure, it can be seen that the performance of EI-IFCM and IFCM algorithms fluctuates with changes in parameters. There are differences in the sensitivity of different datasets to parameters. For the Avila dataset, when $m=2.1$ and $\alpha=0.4$, the RI of the EI-IFCM algorithm achieves good performance. When $m=1.5$ and $\alpha=0.2$, the IFCM algorithm achieves better performance. In this experiment, when $m=3.9$ and $\alpha=1$, EI-IFCM and IFCM achieved the optimal RI values on both the Satimage and Seeds datasets. In addition, for the Abalone dataset, EI-IFCM achieved optimal performance at $m=2.1$ and $\alpha=0.4$, while IFCM achieved optimal performance at $m=3$ and $\alpha=0.7$. When the hesitancy parameter α is 1, the hesitancy of IFS can be obtained by using $\pi_{IFS} = 1 - \mu_A(x)$, which

Fig. 7 Plot of hesitation parameter α on clustering performance of two algorithms

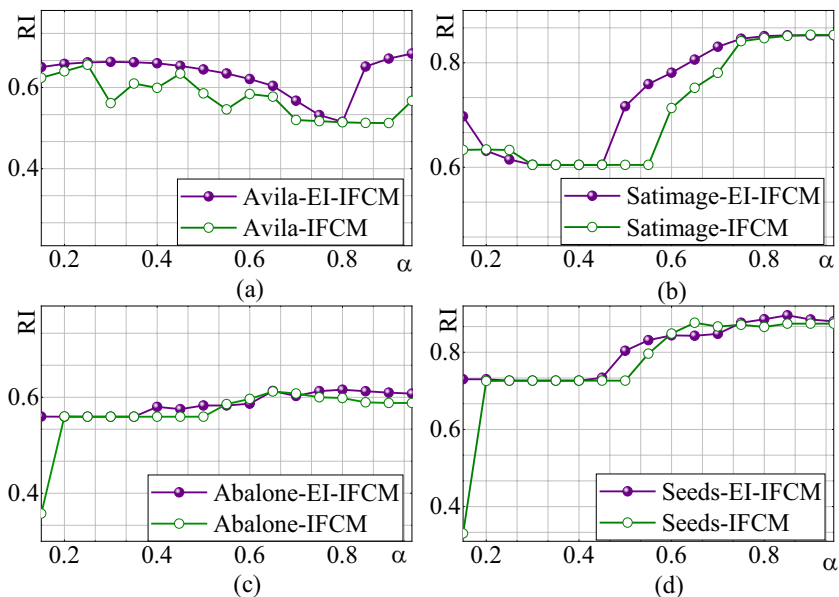
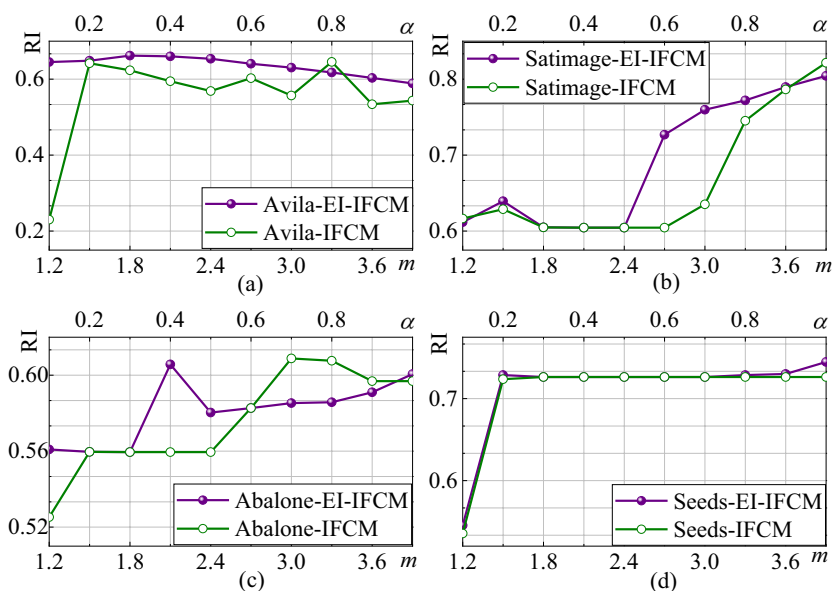


Fig. 8 Plot of m and α on clustering performance of two algorithms



means that the consideration of hesitant information in the data sample is maximized. This means that the consideration of hesitant information in the data sample reaches its maximum state.

From the above analysis, it can be concluded that different datasets and algorithms have different sensitivity to parameters. M represents the fuzzy performance of the dataset, while α represents the hesitant information of the uninterrupted dataset. From the analysis, it can be seen that when both parameters are considered simultaneously, higher clustering performance can be achieved than IFCM on most datasets. Therefore, when solving different practical problems, it is necessary to assign different parameters.

Based on the above analysis, it can be concluded that the proposed algorithm is competitive. Based on the above experiments on algorithms performance and parameter sensitivity analysis, the following conclusions can be obtained. Firstly, the EI-IFCM algorithm is competitive compared to other algorithms. From Table 1, it can be seen that the various indicator values of the proposed algorithm are significantly better than other algorithms. Secondly, compared to other excellent algorithms, the proposed new algorithm uses a data density-based approach to obtain the initial clustering center. Compared with other algorithms that randomly select initial clustering centers, the proposed algorithm starts its learning process from a better center, thereby saving

learning resources. At the same time, it avoids the instability of the initial clustering centers caused by random selection. Thirdly, by using EI condition to segment the dataset. It can assign membership boundaries to data objects that satisfy EI condition, which can avoid the disadvantage that all samples always affecting all clustering centers. Moreover, by applying the proposed algorithm to datasets in different fields, it can be concluded that the algorithm has applicability in fields such as image classification, plant category recognition, medical disease classification, and others. Finally, by comparing the parameter sensitivity analysis between the classical algorithm and the proposed algorithm, it can be seen that the proposed algorithm inherits the characteristics of the classical algorithm with different sensitivities on different datasets. The EI-IFCM algorithm also has different sensitivities on different datasets. This needs further improvement in future research.

5 Conclusion

Due to the idea of IFCM fusing hesitant information into membership and the recently proposed MSFCM model fusing membership boundaries, this paper proposes a new IFCM clustering algorithm called EI-IFCM, which integrates local density. In the EI-IFCM algorithm, the initial clustering centers are obtained based on the sample distribution density and are responsible for reducing instability caused by random selection. Therefore, the proposed algorithm can start its learning process from scratch with relatively stable initial clustering centers. On the other hand, in response to the viewpoint that all samples always affect all clustering centers in the classical IFCM algorithm, the algorithm combines the EI condition and the membership boundary derived in the paper. In this contribution, considering that the EI can provide partition condition for the datasets in two adjacent iterations, different membership calculation rules are given to ensure fast convergence of data samples. To verify the applicability of the proposed algorithm, experiments are conducted on 9 real world datasets. The experimental results show that the proposed EI-IFCM model has a competitive overall performance compared to other advanced models. By conducting experiments on these real world datasets, it can be verified that the proposed algorithm can be applied to fields such as medical disease classification, geographic image recognition, plant and animal differentiation, and others. This demonstrates the applicability of the algorithm.

However, the main limitation of the EI-IFCM algorithm is its inability to adaptively select the optimal parameters. Further research can obtain the optimal computational parameters by using heuristic algorithms such as particle swarm optimization or genetic algorithms. Moreover, this study also

inherits the limitation of classical algorithms that require presetting the number of clusters. Integrating data distribution density into the algorithm to independently obtain the number of clusters can also serve as an improvement to this model, which is one of the future research directions.

Acknowledgment This work was supported by the National Natural Science Foundation of China with grant number of 62173025 and Key Research and Development Project of Guangdong Province with grant number of 2021B0101420003.

Authors contribution Qianxia Ma: Methodology, Writing – original draft, Software, Validation, Formal analysis, Investigation, Visualization, Supervision. Xiaomin Zhu: Resources, Supervision, Project administration. Xiangkun Zhao: Software, Investigation, Writing – review & editing. Butian Zhao: Software, Investigation, Writing – review & editing. Guanhua Fu: Resources, Supervision, Project administration. Runtong Zhang: Resources, Supervision, Project administration.

Data availability Authors confirm that the data supporting the findings of this study are available within the paper.

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Ethical and informed consent for data used Data used in this study obtained informed consent and adhered to ethical guidelines.

Competing interests The authors declare that they have no known competing financial interests or personal relationships.

Reference

1. Zhao F, Liu F, Li C et al (2021) Coarse-fine surrogate model driven multiobjective evolutionary fuzzy clustering algorithm with dual memberships for noisy image segmentation. *Appl Soft Comput* 112:107778. <https://doi.org/10.1016/j.asoc.2021.107778>
2. Wang L (2022) Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization. *Appl Soft Comput* 114:108153. <https://doi.org/10.1016/j.asoc.2021.108153>
3. Baraldi A, Blonda P (1999) A survey of fuzzy clustering algorithms for pattern recognition - Part I. *IEEE Trans Syst Man Cybern Part B-Cybern* 29:778–785. <https://doi.org/10.1109/3477.809032>
4. Sander J, Ester M, Kriegel HP, Xu XW (1998) Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Min Knowl Discov* 2:169–194. <https://doi.org/10.1023/A:1009745219419>
5. Sun G, Cong Y, Dong J et al (2022) What and How: generalized lifelong spectral clustering via dual memory. *IEEE Trans Pattern Anal Mach Intell* 44:3895–3908. <https://doi.org/10.1109/TPAMI.2021.3058852>
6. Ay M, Ozbakir L, Kulluk S et al (2023) FC-Kmeans: Fixed-centered K-means algorithm. *Expert Syst Appl* 211:118656. <https://doi.org/10.1016/j.eswa.2022.118656>
7. Zhao P, Zhang Y, Ma Y et al (2023) Discriminatively embedded fuzzy K-Means clustering with feature selection strategy. *Appl Intell*. <https://doi.org/10.1007/s10489-022-04376-5>

8. Bellman R, Kalaba R, Zadeh L (1966) Abstraction and pattern classification. *J Math Anal Appl* 13:1-. [https://doi.org/10.1016/0022-247X\(66\)90071-0](https://doi.org/10.1016/0022-247X(66)90071-0)
9. Zadeh LA (1965) Fuzzy sets. *Information & Control* 8:338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
10. Yang F, Liu Z, Bai X, Zhang Y (2022) An improved intuitionistic fuzzy C-Means for ship segmentation in infrared images. *IEEE Trans Fuzzy Syst* 30:332–344. <https://doi.org/10.1109/TFUZZ.2020.3037972>
11. Gao Y, Wang Z, Xie J, Pan J (2022) A new robust fuzzy c-means clustering method based on adaptive elastic distance. *Knowledge-Based Syst* 237:107769. <https://doi.org/10.1016/j.knosys.2021.107769>
12. Zhang X, Jian M, Sun Y et al (2020) Improving image segmentation based on patch-weighted distance and fuzzy clustering. *Multimed Tools Appl* 79:633–657. <https://doi.org/10.1007/s11042-019-08041-x>
13. Solanki R, Kumar D (2023) Probabilistic intuitionistic fuzzy c-means algorithm with spatial constraint for human brain MRI segmentation. *Multimed Tools Appl* 82:33663–33692. <https://doi.org/10.1007/s11042-023-14512-z>
14. Ma Q-X, Zhu X-M, Bai K-Y, et al A novel failure mode and effect analysis method with spherical fuzzy entropy and spherical fuzzy weight correlation coefficient. 21
15. Varshney AK, Muhuri PK, Lohani QMD (2022) Density-based IFCM along with its interval valued and probabilistic extensions, and a review of intuitionistic fuzzy clustering methods. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-022-10236-y>
16. Atanassov K (1986) Intuitionistic fuzzy-sets. *Fuzzy Sets Syst* 20:87–96. [https://doi.org/10.1016/S0165-0114\(86\)80034-3](https://doi.org/10.1016/S0165-0114(86)80034-3)
17. Hazarika BB, Gupta D, Borah P (2021) An intuitionistic fuzzy kernel ridge regression classifier for binary classification. *Appl Soft Comput* 112:107816. <https://doi.org/10.1016/j.asoc.2021.107816>
18. Çakır E, Taş MA (2023) Circular intuitionistic fuzzy decision making and its application. *Expert Syst Appl* 225:120076. <https://doi.org/10.1016/j.eswa.2023.120076>
19. Mishra U, Gupta D, Hazarika BB (2023) An intuitionistic fuzzy random vector functional link classifier. *Neural Process Lett* 55:4325–4346. <https://doi.org/10.1007/s11063-022-11043-w>
20. Hazarika BB, Gupta D, Gupta U (2023) Intuitionistic fuzzy kernel random vector functional link classifier. In: Sisodia DS, Garg L, Pachori RB, Tanveer M (eds) *Machine Intelligence Techniques for Data Analysis and Signal Processing*. Springer Nature Singapore, Singapore, pp 881–889
21. Hajek P, Froelich W, Prochazka O (2020) Intuitionistic fuzzy grey cognitive maps for forecasting interval-valued time series. *Neurocomputing* 400:173–185. <https://doi.org/10.1016/j.neucom.2020.03.013>
22. Xu Z, Wu J (2010) Intuitionistic fuzzy C-means clustering algorithms. *J. Syst Eng Electron* 21:580–590. <https://doi.org/10.3969/j.issn.1004-4132.2010.04.009>
23. Chaira T (2011) A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. *Applied Soft Computing* 11:1711–1717. <https://doi.org/10.1016/j.asoc.2010.05.005>
24. Singh C, Ranade SK, Kaur D, Bala A (2024) A novel approach for brain MRI segmentation and image restoration under intensity inhomogeneity and noisy conditions. *Biomedical Signal Processing and Control* 87:105348. <https://doi.org/10.1016/j.bspc.2023.105348>
25. Liu C, Li Z, Wu Z et al (2023) An unsupervised snow segmentation approach based on dual-polarized scattering mechanism and deep neural network. *IEEE Trans Geosci Remote Sensing* 61:1–14. <https://doi.org/10.1109/TGRS.2023.3262727>
26. Varshney AK, Muhuri PK, Danish Lohani QM (2022) PIFHC: The probabilistic intuitionistic fuzzy hierarchical clustering algorithm. *Applied Soft Computing* 120:108584. <https://doi.org/10.1016/j.asoc.2022.108584>
27. Cherif S, Alimi AM (2022) Novel Intuitionistic-Based Interval Type-2 Fuzzy Similarity Measures With Application to Clustering. *IEEE TRANSACTIONS ON FUZZY SYSTEMS* 30
28. Wu C, Huang C, Zhang J (2023) Local information-driven intuitionistic fuzzy c-Means algorithm integrating total bregman divergence and kernel metric for noisy image segmentation. *Circuits Syst Signal Proc* 42:1522–1572. <https://doi.org/10.1007/s00034-022-02175-4>
29. Zhou S, Li D, Zhang Z, Ping R (2021) A new membership scaling fuzzy C-Means clustering algorithm. *IEEE Trans Fuzzy Syst* 29:2810–2818. <https://doi.org/10.1109/TFUZZ.2020.3003441>
30. Yang M-S, Nataliani Y (2018) A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy. *IEEE Trans Fuzzy Syst* 26:817–835. <https://doi.org/10.1109/TFUZZ.2017.2692203>
31. Hathaway RJ, Hu Y (2009) Density-weighted fuzzy c-Means clustering. *IEEE Trans Fuzzy Syst* 17:243–252. <https://doi.org/10.1109/TFUZZ.2008.2009458>
32. Li J, Endo Y (2020) Fuzzy c-Means with improved particle swarm optimization. In: *2020 Ieee International Conference on Fuzzy Systems (fuzz-IEEE)*. Ieee, New York, p 22212
33. Zhu X, Wu X, Wu B, Zhou H (2023) An improved fuzzy C-means clustering algorithm using Euclidean distance function. *IFS* 44:9847–9862. <https://doi.org/10.3233/JIFS-223576>
34. Hashemi SE, Gholian-Jouybari F, Hajiaghahi-Keshteli M (2023) A fuzzy C-means algorithm for optimizing data clustering. *Expert Syst Appl* 227:120377. <https://doi.org/10.1016/j.eswa.2023.120377>
35. Joaquín Pérez, Pires CE, Balby L, et al (2013) Early classification: a new heuristic to improve the classification step of K-Means. *J Inform Data Manage*. <https://doi.org/10.5772/intechopen.85447>
36. Bustince H, Kacprzyk J, Moledano V (2000) Intuitionistic fuzzy generators - Application to intuitionistic fuzzy complementation. *Fuzzy Sets Syst* 114:485–504. [https://doi.org/10.1016/S0165-0114\(98\)00279-6](https://doi.org/10.1016/S0165-0114(98)00279-6)
37. Pei H-X, Zheng Z-R, Wang C et al (2017) D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation. In: Ahuja V, Shi Y, Khazanchi D et al (eds) *5th International Conference on Information Technology and Quantitative Management*. Itqm 2017. Elsevier Science Bv, Amsterdam, pp 407–414
38. Havens TC, Bezdek JC, Leckie C et al (2012) Fuzzy c-Means algorithms for very large data. *IEEE Trans Fuzzy Syst* 20:1130–1146. <https://doi.org/10.1109/TFUZZ.2012.2201485>
39. Chen Y, Zhou S, Zhang X et al (2022) Improved fuzzy c -means clustering by varying the fuzziness parameter. *Pattern Recog Lett* 157:60–66. <https://doi.org/10.1016/j.patrec.2022.03.017>
40. Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Ame Statist Association*. <https://doi.org/10.1080/01621459.1971.10482356>
41. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2:193–218. <https://doi.org/10.1007/BF01908075>
42. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J Mach Learn Res* 11:2837–2854
43. Fowlkes EB, Mallows CL (1983) A Method for Comparing Two Hierarchical Clusterings: Comment. *Publi Ame Statistic Assoc* 78:553–569. <https://doi.org/10.1080/01621459.1983.10478008>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Qianxia Ma received the B.S. degree from Lanzhou University of Technology. She is currently studying for his doctorate degree in the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, China. Her current research interests include machine learning and fault diagnosis.



Guanhua Fu received the M.S. degree from Xiangtan University. He is currently working in RailTransit Department, Tianjin Jinhang Computing Technology Research Institute, China. His current research interests include machine learning and fault diagnosis.



Xiaomin Zhu received the Ph.D. degree from the Technical University of Crete in Greece in 2004, and received her B.S. and M.S. degrees from Tianjin University, Tianjin, China, in 1985 and in 1988, respectively. She is presently a professor with the Department of Industrial Engineering at Beijing Jiaotong University. Her research interests include system evaluation and optimization, modeling and simulation, optimal control, operations research. She is a PI

for over 60 research projects, and has published 80 papers and 15 books in these fields.



Runtong Zhang received the Ph.D. degree in Production Engineering and Management from Technical University of Crete in Greece in 1996, and his B.S. in Computer Science and Automation from the Dalian Maritime University in China in 1985, respectively. He is presently a professor and the head of Department of Information Management at Beijing Jiaotong University, China. He was also with the Swedish Institute of Computer Science as a senior researcher, and

the Port of Tianjin Authority as an engineer. His current research interests include big data driven decision and management operations research and artificial intelligence. He has published over 400 papers in referenced journals and conferences, and 40 books. He has been a PI for over 100 research projects and is a holder of 15 patents. He has been Senior Member of IEEE, and a general chair or co-chair for over 10 IEEE sponsored international conferences.



Xiangkun Zhao received the B.S. degree from Beijing Jiaotong Univc. He is currently studying for his doctorate degree in the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, China. His current research interests include machine learning and fault diagnosis.



Butian Zhao received the B.S. degree from Beijing Jiaotong University. He is currently studying for his doctorate degree in the School of Economics and Management, Beijing Jiaotong University, China. His current research interests include machine learning and health informatics.