International Journal of Machine Learning and Cybernetics (2021) 12:859–875
https://doi.org/10.1007/s13042-020-01206-3

ORIGINAL ARTICLE

Intuitionistic fuzzy c-means clustering algorithm based on a novel weighted proximity measure and genetic algorithm

Wen-hui Hou[1] · Yi-ting Wang[1] · Jian-qiang Wang[1,2] · Peng-Fei Cheng[2] · Lin Li[3]

Received: 4 September 2019 / Accepted: 18 September 2020 / Published online: 9 October 2020

Abstract

In the era of big data, the research on clustering technologies is a popular topic because they can discover the structure of complex data sets with minimal prior knowledge. Among the existing soft clustering technologies, as an extension of fuzzy c-means (FCM) algorithm, the intuitionistic FCM (IFCM) algorithm has been widely used due to its superiority in reducing the effects of outliers/noise and improving the clustering accuracy. In the existing IFCM algorithm, the measurement of proximity degree between a pair of objects and the determination of parameters are two critical problems, which have considerable effects on the clustering results. Therefore, we propose an improved IFCM clustering technique in this paper. Firstly, a novel weighted proximity measure, which aggregates weighted similarity and correlation measures, is proposed to evaluate not only the closeness degree but also the linear relationship between two objects. Subsequently, genetic algorithms are utilized for identifying the optimal parameters. Lastly, experiments on the proposed IFCM technique are conducted on synthetic and UCI data sets. Comparisons with other approaches in cluster evaluation indexes indicate the effectiveness and superiority of our method.

Keywords Intuitionistic fuzzy c-means (IFCM) algorithm · Intuitionistic fuzzification · Similarity measure · Correlation coefficient · Genetic algorithm (GA)

1 Introduction

As an extension of fuzzy set (FS) [1], intuitionistic FS (IFS), which consist of membership, non-membership and hesitation degrees, were incepted in [2] to describe and process data with uncertainty. IFS has been continuously studied and applied to various fields, such as pattern recognition, image processing, decision making and clustering [3]. Out of all the applications, the clustering techniques of IFS are among the major domains that have been found to be highly useful but rarely investigated.

Clustering refers to an exploratory data analysis tool for discovering the data structure in multivariate data sets through association rules and grouping all data into multiple clusters. A good clustering result requires that the items within the same cluster have a maximal degree of association and minimal otherwise. With the arrival of the big data age, many realistic problems concerning taxonomical, geological, medical, engineering and business systems require clustering techniques to solve. Thus, the development of clustering technologies is gaining increasing attention from researchers. Traditional clustering techniques can be broadly classified into hard and soft clustering. Hard clustering algorithms assume that a clear boundary exists among different clusters and assign each object to a single cluster exactly. However, in many real-world applications, a sharp boundary does not necessarily exist among clusters; an object may belong to multiple clusters. For this reason, many soft clustering algorithms have been studied. Lingras et al. [4] introduced rough set theory into clustering and proposed the rough k-means clustering algorithm, which assigns objects to multiple clusters in accordance with the concept of upper and lower approximation of rough sets. Based on FS theory,

✉ Jian-qiang Wang
jqwang@csu.edu.cn

[1] School of Business, Central South University, Changsha 410083, People's Republic of China

[2] Hunan Engineering Research Center for Intelligent Decision Making and Big Data on Industrial Development, Hunan University of Science and Technology, Xiangtan 411201, People's Republic of China

[3] School of Business, Hunan University, Changsha 410082, People's Republic of China

the fuzzy c-means (FCM) clustering algorithm, which allows each data to be subordinate to multiple clusters with varying membership degree that represents the proximity of the data to multiple cluster centres, has been developed [5]. As a conceptual bridge between rough set and FS, shadow set has also been successfully applied to clustering; for example, shadowed c-means [6] and k-means [7] clustering algorithms have been introduced. Inspired by three-way decision, Yu et al. [8, 9] proposed three-way clustering theory, which divides the entire area into three parts, namely, positive, negative and boundary areas, to represent the three states of an object: belong to, not belong to and uncertain.

Among several well-known soft clustering techniques, FCM algorithm is widely used in various fields due to its high efficiency and ease of use. However, the main shortcoming of this algorithm is that it is sensitive to noise and outliers [10]. IFSs can express more valuable information than crisp data sets, and using IFSs to represent a project may reduce the effect of noise/outliers on clustering results. Thus, the intuitionistic FCM (IFCM) clustering algorithm, which adopts the weighted Euclidean distance between IFSs in the objective function of FCM algorithm to cope with uncertainty, was firstly introduced in [11]. Many researchers have then used this IFCM algorithm to solve various problems, such as image segmentation [12], geodemographic analysis [13], customer segmentation [14] and forecasting [15]. All of these studies have concluded that compared with the conventional FCM, the IFCM algorithm can be less affected by noise, produces more accurate clustering results and requires fewer iterations. Nevertheless, two critical issues occur in the IFCM algorithm, and the specific analyses are as follows.

One critical issue in the IFCM algorithm is the proximity measurement among objects. In many studies, the distance/similarity measures between IFSs have been adopted to determine the proximity degree between a pair of items. Xu's IFCM algorithm uses the Euclidean distance between IFSs as a proximity measure. However, the Euclidean distance often results in low clustering accuracy when noise exists in data sets [16]. Therefore, many scholars have dedicated efforts to find a suitable proximity measure for cluster analysis. In [17], the author compared the performance of several IFCM algorithms with different intuitionistic fuzzy similarity measures on UCI data sets. In [18], several well-known intuitionistic fuzzy distance metrics were reviewed and experimented on a synthetic data set and real data sets. Moreover, many new distance/similarity measures of IFSs for clustering techniques have been introduced in recent studies [19–21]. In accordance with the experimental results on various data sets, all of the above-mentioned clustering algorithms with different new intuitionistic fuzzy distance/similarity measures have been proved to perform better than clustering algorithms with traditional Euclidean distance. However, the use of distance/similarity measures to find the proximity degree between a pair of objects only evaluates the closeness degree between the two projects and ignores the linear relationship between them. That is, it cannot identify the correlations between projects with similar trends. In this case, some valuable information will be lost during the clustering process, which will affect the precision of clustering results.

The other critical issue is the determination of parameters. In the IFCM algorithm, the user should set many parameters in advance, and suitable parameters can promote the success of the clustering algorithm [22]. In the conventional algorithm, these parameters are often subjectively set, but this condition may lead to unreasonable clustering results due to problem complexity and the lack of knowledge of the user. Therefore, many studies have adopted several optimization algorithms, such as grid search, stochastic gradient descent and adaptive approach, to find optimal parameters objectively [20]. The experimental results show that determining parameters with objective methods is more robust than that with subjective methods. Nonetheless, the performance of the optimization algorithms mentioned above remains unideal in terms of calculation speed and accuracy, and they will not assist the IFCM algorithm in producing the best clustering result.

After clearly recognizing that the aforementioned two key issues will weaken the clustering effect of the IFCM algorithm, this study aims to find a suitable method to improve the existing IFCM algorithm from the two aspects.

For the first issue, we consider correlation measure as one of the most widely used metrics that can compensate for the defect of distance/similarity measures. In statistical analysis, the correlation coefficient, which evaluates the strength and direction of the linear relationship between two sets of data, can consider the trend of each set of data. The method of correlation measurement in various fuzzy environments has been greatly developed and applied in many fields. The concept of correlation among IFSs was firstly proposed by [23] to measure the interrelation of IFSs. Later, various correlation measures of IFSs [24–27] were developed and widely used in clustering. The superiority of combining similarity and correlation measures for cluster analysis has been illustrated in [28].

For the second issue, many studies have proved that heuristic optimization algorithms can perform better than other optimization algorithms (such as grid search, random search and stochastic gradient descent), especially on complex issues [29]. Thus, we believe that heuristic optimization algorithms have remarkable advantages in terms of objectively determining parameters. Several heuristic optimization algorithms, such as genetic algorithm (GA) [30], particle swarm optimization (PSO) [31] and artificial bee colony (ABC) [14], have been widely used in many studies on account of their excellent performance in dealing with highly complicated problems. Among these optimization algorithms, GA is a random search algorithm that simulates the biological evolutionary process and is applied most frequently because of its proven powerful global search capabilities. In [32, 33], the authors adopted GA to select the optimal parameters of their extended IFCM algorithm, and the results demonstrated the efficiency of GA in improving the performance of clustering algorithms.

Inspired by the preceding analysis, we propose a new IFCM algorithm which improves the existing IFCM algorithm from two aspects: proximity measurement among objects and parameter determination.

Firstly, we propose a new weighted proximity measure for the IFCM algorithm, which aggregates weighted similarity and correlation measures to evaluate not only the closeness degree but also the linear relationship between two objects represented by IFSs. The weight of each feature in the weighted proximity measure is calculated using the maximizing deviation method to reduce information loss. Then, GA is applied to determine the optimal parameters of this clustering algorithm for avoiding the adverse effects of subjectively setting parameters on clustering results. The time complexity of the improved GA-IFCM algorithm is also analysed. Lastly, we conduct three experiments on a synthetic data set and five UCI data sets to compare the improved GA-IFCM algorithm based on the proposed weighted proximity measure and parameter optimization algorithms with other clustering methods in terms of accuracy rate, four validation indexes and time consumption.

The remainder of this paper is organized as follows. Section 2 presents some basic concepts related to IFS and the process of the conventional IFCM algorithm. Section 3 introduces an improved IFCM algorithm based on the proposed novel weighted proximity measure and GA. Section 4 discusses the experimental results of the proposed clustering technique on synthetic and UCI data sets. Lastly, the conclusions and future research directions are stated in Sect 5.

2 Preliminaries

In this section, we introduce some basic concepts related to IFS and the intuitionistic fuzzy clustering approach, which will be utilized hereinafter.

2.1 IFS

Definition 1 IFS [2]. An IFS A in a universe X can be written as $A=\left\{\langle x,\mu_A(x),v_A(x)\rangle | x\in X\right\}$, where $\mu_A(x):X\rightarrow[0,1]$ and $v_A(x):X\rightarrow[0,1][0,1]$, with the condition $0\leq\mu_A(x)+v_A(x)\leq 1$, represent the membership degree and non-membership degree of the element x to A. The hesitation degree of x to A is expressed as $\pi_A(x)=1-\mu_A(x)-v_A(x)$, evidently, $0\leq\pi_A(x)\leq 1, x\in X$.

2.1.1 Distance/similarity measures between IFSs

Definition 2 Distance measure between IFSs [34]. Let A and B be two IFSs on X, mapping $D:IFS\times IFS\rightarrow[0,1]$; $D(A,B)$ represents the distance between A and B, and it needs to satisfy the following properties:

$(P1)\, 0\leq D(A,B)\leq 1$

$(P2)\, D(A,B)=0\ \textit{if and only if}\ A=B$

$(P3)\, D(A,B)=D(B,A)$

$(P4)\, \textit{If}\ A\subseteq B\subseteq C\ \textit{then}\ D(A,C)\geq D(A,B)$

Definition 3 Similarity measure between IFSs [34]. Let A and B be two IFSs on X, mapping $S:IFS\times IFS\rightarrow[0,1]$; $S(A,B)$ represents the similarity between A and B, and it needs to satisfy the following properties:

$(P1)\, 0\leq S(A,B)\leq 1$

$(P2)\, S(A,B)=1\ \textit{if and only if}\ A=B$

$(P3)\, S(A,B)=S(B,A)$

$(P4)\, \textit{If}\ A\subseteq B\subseteq C\ \textit{then}\ S(A,C)\leq S(A,B)$

The similarity measure is the complementary concept of the distance measure. Therefore, if $D(A,B)$ denotes the distance between A and B, then $S(A,B)=1-D(A,B)$ is the similarity measure between A and B.

Similarity and distance measurements among data points are important components of clustering algorithms. Many of them involving IFS have been proposed in previous literature. The weight of each element $x_i\in X$ contains important information. Therefore, several well-known weighted similarity and distance measures that will be used later for comparative analysis are shown below.

We set $\omega = \left(\omega_1, \omega_2, \ldots, \omega_n\right)$ as the weight vector of $x_i (i = 1, 2, \ldots, n)$, with $\omega_i \geq 0 (i = 1, 2, \ldots, n)$, and $\sum_{i=1}^{n} \omega_i = 1$.

The weighted Hamming distance of A and B [35] is

$$D_{\omega Ham}(A,B) = \frac{1}{2} \sum_{i=1}^{n} \omega_i \left(\left| \mu_A(x_i) - \mu_B(x_i) \right| + \left| \nu_A(x_i) - \nu_B(x_i) \right| + \left| \pi_A(x_i) - \pi_B(x_i) \right| \right) \tag{1}$$

The weighted Euclidean distance of A and B [35] is

$$D_{\omega E}(A,B) = \sqrt{\frac{1}{2} \sum_{i=1}^{n} \omega_i \left(\left(\mu_A(x_i) - \mu_B(x_i) \right)^2 + \left(\nu_A(x_i) - \nu_B(x_i) \right)^2 + \left(\pi_A(x_i) - \pi_B(x_i) \right)^2 \right)} \tag{2}$$

The weighted cosine similarity of A and B [36] is

$$S_{\omega Cos}(A,B) = \sum_{i=1}^{n} \omega_i \frac{\mu_A(x_i)\mu_B(x_i) + \nu_A(x_i)\nu_B(x_i)}{\sqrt{\left(\mu_A(x_i)\right)^2 + \left(\nu_A(x_i)\right)^2} \sqrt{\left(\mu_B(x_i)\right)^2 + \left(\nu_B(x_i)\right)^2}} \tag{3}$$

The weighted tangent similarity of A and B [37] is

$$S_{\omega Tan}(A,B) = \sum_{i=1}^{n} \omega_i \left\langle 1 - \tan\left(\frac{\pi\left(\left| \mu_A(x_i) - \mu_B(x_i) \right| + \left| \nu_A(x_i) - \nu_B(x_i) \right| + \left| \pi_A(x_i) - \pi_B(x_i) \right| \right)}{12} \right) \right\rangle \tag{4}$$

These well-known similarity and distance measures have many shortcomings, such as the possibility of generating counter-intuitive results, zero division, graphical interpretation and intelligibility. To solve these problems, a large number of studies on the distance/similarity metrics of IFSs have emerged in recent years. Amongst them, a new distance measure amongst IFSs based on transformed isosceles triangles, which was proposed by [34], can represent higher discernibility than most existing methods. It can be expressed as

$$D_{\omega Jiang}(A,B) = \frac{1}{2} \sum_{i=1}^{n} \omega_i \left(\begin{array}{l} \left| \frac{2\left(\mu_A(x_i)\pi_B(x_i) - \mu_B(x_i)\pi_A(x_i)\right) - 4\left(\mu_A(x_i) - \mu_B(x_i)\right)}{4 - \pi_A(x_i)\pi_B(x_i)} \right| \\ + \left| \frac{4\left(\nu_A(x_i) - \nu_B(x_i)\right) + 2\left(\nu_A(x_i)\pi_B(x_i) - \nu_B(x_i)\pi_A(x_i)\right) + 2\left(\pi_A(x_i) - \pi_B(x_i)\right)}{4 - \pi_A(x_i)\pi_B(x_i)} \right| \end{array} \right) \tag{5}$$

2.1.2 Correlation measure of IFSs

Definition 4 Correlation measure of IFSs [26]. Let A and B be two IFSs on X, mapping $C : IFS \times IFS \rightarrow [0, 1]; C(A, B)$ represents the correlation measure between A and B, and it needs to satisfy the following properties:

$(P1)\, 0 \leq C(A,B) \leq 1$

$(P2)\, C(A,B) = 1$ *if and only if* $A = B$

$(P3)\, C(A,B) = C(B,A)$

Since the twenty-first century, many studies [24, 25, 27] have introduced various methods to calculate the coefficient of IFSs. As one of the several well-known correlation coefficients of IFSs, the method proposed in [26] has been used in clustering algorithms. It considers membership degree, non-membership degree, hesitation degree and the weight of element x_i. So it retains more information than other correlation coefficients. The specific expression is as follows:

$$\rho_{\omega Xu}(A,B) = \frac{C_\omega(A,B)}{\max\left(T_\omega(A), T_\omega(B)\right)}, \tag{6}$$

where

$$C_\omega(A,B) = \sum_{i=1}^{n} \omega_i \left[\mu_A(x_i)\mu_B(x_i) + \nu_A(x_i)\nu_B(x_i) + \pi_A(x_i)\pi_B(x_i) \right]. \tag{7}$$

$T_{\omega}(A) = C_{\omega}(A, A)$ and $T_{\omega}(B) = C_{\omega}(B, B)$. The values of the correlation coefficients change within the interval [0, 1].

2.2 Intuitionistic fuzzy clustering approach

The proposed intuitionistic fuzzy clustering approach involves two stages. The first one is to map crisp values to IFSs through the intuitionistic fuzzification method, and the second one is to cluster IFSs by using the IFCM algorithm.

2.2.1 Intuitionistic fuzzification

With reference to [17], in the process of intuitionistic fuzzification, the crisp data set needs to be converted into a fuzzy data set firstly and be converted into an intuitionistic fuzzy data set subsequently.

Let X be the data set with n objects, and each object consists of d features. The fuzzy membership function of each data element x_{ij} is defined as

$$\mu(x_{ij}) = \frac{x_{ij} - \min(x_{ij})}{\max\left(x_j\right) - \min(x_j)}, \tag{8}$$

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

The IFS membership function is equal to fuzzy membership function $\mu\left(x_{ij}\right)$. IFS non-membership function $v\left(x_{ij}\right)$ can be calculated from Yager's generating function [38].

$$N(x) = (1 - x^{\alpha})^{\frac{1}{\alpha}}, \quad \alpha > 0, \tag{9}$$

where $N(1) = 0, N(0) = 1$.

Therefore, IFS A can be written as

$$A = \left\{ \left\langle x_{ij}, \mu\left(x_{ij}\right), \left(1 - \mu\left(x_{ij}\right)^{\alpha}\right)^{\frac{1}{\alpha}} \right\rangle \middle| x_{ij} \in X \right\}. \tag{10}$$

The hesitation degree is

$$\pi\left(x_{ij}\right) = 1 - \mu\left(x_{ij}\right) - \left(1 - \mu\left(x_{ij}\right)^{\alpha}\right)^{\frac{1}{\alpha}}. \tag{11}$$

The value of α is used to control the hesitation degree of IFSs. When parameter α is equal to $1, v\left(x_{ij}\right) = 1 - \mu\left(x_{ij}\right)$, which means that the hesitation degree is 0 and is not considered.

2.2.2 IFCM clustering algorithm

The FCM clustering algorithm partitions a set of data into c clusters on the basis of Euclidean distance metrics. As an improvement of the FCM algorithm, the IFCM algorithm developed by [11] replaces the traditional Euclidean distance of crisp data with intuitionistic fuzzy similarity/distance measures.

Given n IFSs $\left\{A_1, A_2, \ldots, A_n\right\}$ are to be divided into $c(1 \le c \le n)$ clusters. The method aims at minimizing the objective function, that is,

$$\begin{aligned} &\min J_{IFCM}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(u_{ij}\right)^m d^2\left(A_i, V_j\right) \\ &s.t \sum_{j=1}^{c} u_{ij} = 1, \quad u_{ij} \in [0, 1], \quad 0 \le \sum_{i=1}^{n} u_{ij} \le n, \end{aligned} \tag{12}$$

where d is the Euclidean distance between object A_i and the jth cluster centre V_j; u_{ij} is the membership degree of A_i belonging to the jth cluster; parameter $m(m > 1)$ is the weighting exponent that determines the fuzziness of the clustering result, and the default value of m is 2.

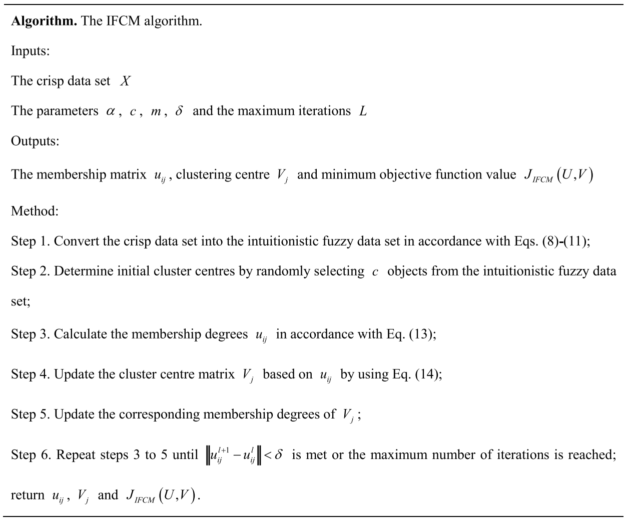
The Lagrange multiplier method is used to minimize Eq. (12), and the updated equations of u_{ij} and V_j are given as follows:

$$u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d(A_i, V_j)}{d(A_i, V_l)} \right)^{\frac{1}{1-m}}} \tag{13}$$

$$V_j = \frac{\sum_{i=1}^{n} \left(u_{ij}\right)^m A_i}{\sum_{i=1}^{n} \left(u_{ij}\right)^m}. \tag{14}$$

The IFCM algorithm optimizes J_{IFCM} by continuously updating u_{ij} and V_j until $\left\| u_{ij}^{l+1} - u_{ij}^{l} \right\| < \delta$ is satisfied, where δ is the termination tolerance for solution accuracy.

The specific procedure of the IFCM algorithm can be stated as follows.



```
**Algorithm.** The IFCM algorithm.
Inputs:
The crisp data set $X$
The parameters $\alpha$, $c$, $m$, $\delta$ and the maximum iterations $L$
Outputs:
The membership matrix $u_{ij}$, clustering centre $V_j$ and minimum objective function value $J_{IFCM}(U,V)$
Method:
Step 1. Convert the crisp data set into the intuitionistic fuzzy data set in accordance with Eqs. (8)-(11);
Step 2. Determine initial cluster centres by randomly selecting $c$ objects from the intuitionistic fuzzy data set;
Step 3. Calculate the membership degrees $u_{ij}$ in accordance with Eq. (13);
Step 4. Update the cluster centre matrix $V_j$ based on $u_{ij}$ by using Eq. (14);
Step 5. Update the corresponding membership degrees of $V_j$;
Step 6. Repeat steps 3 to 5 until $\|u_{ij}^{l+1} - u_{ij}^{l}\| < \delta$ is met or the maximum number of iterations is reached;
return $u_{ij}$, $V_j$ and $J_{IFCM}(U,V)$.
```

This conventional IFCM algorithm has two main disadvantages.

Firstly, the weighted Euclidean distance of IFSs is adopted as the proximity measurement between two objects. However, the Euclidean distance can only evaluate the distance between two projects but ignores the effects of the direction and linear relationship between them. In this case, some valuable information will be lost during the clustering process, which will affect the precision of clustering results.

Secondly, the parameters in this algorithm are subjectively set by the user. These preset parameters are not optimal parameters due to the lack of user knowledge, which will also affect the final clustering results.

To compensate for the two shortcomings mentioned above, this paper proposes new proximity measurement and parameter determination methods and apply them to the IFCM algorithm in the next section.

3 Proposed IFCM algorithm based on the aggregated weighted proximity measure and GA (GA-IFCM)

In this section, our proposed GA-IFCM algorithm is explained. To formalize the GA-IFCM clustering algorithm, we firstly propose a new proximity function $\overline{s}_{\omega}(A, B)$, which aggregates similarity and correlation measures, and then use GA to determine the parameters in the clustering process.

3.1 Aggregated weighted proximity measure for IFSs

In this part, the new proximity measure of IFSs that aggregates similarity and correlation measures is presented.

The proximity measure between two projects is an important component of the IFCM algorithm. As mentioned in Sect. 1, in the existing method, the similarity/distance between two IFSs are used to find proximity between any pair of objects. It is obtained by weighted averaging of the similarity/distance between each pair of intuitionistic fuzzy numbers (IFNs), which ignores the trends of a set of IFNs in an IFS. Thus, the similarity/distance measures cannot identify the correlations between two IFSs with similar trends. On the contrary, in the correlation measure, correlation coefficients can evaluate the strength and direction of the linear relationship between two IFSs. The trends of a set of IFNs in an IFS are considered. Thus, we aggregate similarity measures and correlation coefficients into a new proximity measure of IFSs and then apply it to the IFCM algorithm.

With regard to similarity measures, as shown in Sect. 2.1.1, several well-known similarity/distance measures have been listed. Among them, the Hamming and Euclidean distances measure the similarity between two IFNs on the basis of their distance, ignoring the effect of their direction. By contrast, the cosine and tangent similarities only focus on the direction, neglecting the distance between two IFNs. In fact, the distance can quantify how close two IFNs are from each other, and the direction can quantify how orthogonal they are. Therefore, considering that using the Euclidean distance can produce more accurate clustering results than those by using the Hamming distance for most data sets [18] and that cosine similarity does not satisfy property $P2$ in Definition 3 [36], we combine the weighted Euclidean distance and tangent similarity to define a new similarity measure for IFSs. It can be expressed as follows:

$$S_{\omega}(A, B) = \frac{1}{2}\left(S_{\omega\tan}(A, B) + 1 - D_{\omega E}(A, B)\right) \tag{15}$$

Measure $S_{\omega}(A, B)$ satisfies the properties of similarity measures described in Definition 3.

With regard to similarity measures, as shown in Sect. 2.1.2, the well-known correlation coefficients $\rho_{\omega Xu}(A, B)$ has been used in clustering algorithms. Therefore, we intend to aggregate it with the similarity measure proposed above as a new proximity measure for clustering.

From [28], the values of $S_{\omega}(A, B)$ and $\rho_{\omega Xu}(A, B)$ are between 0 and 1. Let mapping $f : [0, 1]^n \to [0, 1]$ be idempotent and monotonically increasing, then $f\left(S_{\omega}(A, B), \rho_{\omega Xu}(A, B)\right)$ can also satisfy the common properties of $S_{\omega}(A, B)$ and $\rho_{\omega Xu}(A, B)$ in Definition 4. In this study, we use the most common aggregation operator, the arithmetic averaging operator, as aggregation function f to aggregate similarity measures $S_{\omega}(A, B)$ and correlation coefficients $\rho_{\omega Xu}(A, B)$ into new proximity measure $\overline{s}_{\omega}(A, B)$, that is,

$$\overline{s}_{\omega}(A, B) = \frac{1}{2}\left(S_{\omega}(A, B) + \rho_{\omega Xu}(A, B)\right) \tag{16}$$

Measure $\overline{s}_{\omega}(A, B)$ is not a geometrical measure but an information measure.

Clearly, this proximity measure will aggregate different and conflicting consequences obtained by similarity measures and correlation coefficients. Thus, the corresponding clustering result obtained using the proximity measure is relatively objective, comprehensive and reasonable.

3.2 Computing optimal weight with the maximizing deviation method

In the process of calculating the proximity between two IFSs, the determination of the weight of each feature is an important issue. For data sets with completely unknown weight information, we use the maximizing deviation method proposed in [39] to determine the weight of each feature.

This approach indicates that if the performance value of each object under a certain feature is inconsiderably different, then the value of the useful information provided by the feature is small, and a small weight should be assigned.

Let data set $X = \left\{X_1, X_2, \dots, X_n\right\}$, $X_1, X_2, \dots, X_n$ be n objects, $C_1, C_2, \dots, C_d$ be d features and $x_{ij}(i = 1, 2, \dots, n; j = 1, 2, \dots, d)$ be the performance values of objects $X_i(i = 1, 2, \dots, n)$ under feature $C_j(j = 1, 2, \dots, d)$. The weight of the jth feature is represented by ω_j $\left(\sum_{j=1}^{d} \omega_j = 1, 0 \le \omega_j \le 1\right)$.

For feature C_j, the deviation of object X_i to all other objects can be expressed as

$$W_{ij}(\omega) = \sum_{k=1}^{n} \left|x_{ij} - x_{kj}\right| \omega_j, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, d. \tag{17}$$

Then, the deviations of all objects to other objects under feature C_j are

$$W_j(\omega) = \sum_{i=1}^{n} W_{ij}(\omega) = \sum_{i=1}^{n} \sum_{k=1}^{n} \left|x_{ij} - x_{kj}\right| \omega_j, \quad j = 1, 2, \dots, d. \tag{18}$$

The deviations of all objects for all features should be described as $W(\omega)$. The optimal weight vector $\omega = \left(\omega_1, \omega_2, \dots, \omega_d\right)$ should maximize $W(\omega)$, as shown in the nonlinear programming model below.

$$\begin{aligned} &\max W(\omega) = \sum_{j=1}^{d} W_j(\omega) = \sum_{j=1}^{d} \sum_{i=1}^{n} \sum_{k=1}^{n} \left|x_{ij} - x_{kj}\right| \omega_j \\ &s.t. 0 \le \omega_j, \sum_{j=1}^{d} \omega_j = 1, \quad j = 1, 2, \dots, d. \end{aligned} \tag{19}$$

Model (19) is solved using the Lagrange function, and we obtain a normalizing optimal weight of feature C_j, that is,

$$\omega_j^* = \frac{\sum_{i=1}^{n} \sum_{k=1}^{n} \left|x_{ij} - x_{kj}\right|}{\sum_{j=1}^{d} \sum_{i=1}^{n} \sum_{k=1}^{n} \left|x_{ij} - x_{kj}\right|}, \quad j = 1, 2, \dots, d \tag{20}$$

3.3 GA-IFCM based on the aggregated weighted proximity measure

After the intuitionistic fuzzification presented in Sect. 2.2.1, the data set $X = \left\{X_1, X_2, \dots, X_n\right\}$ becomes intuitionistic fuzzy data set $A = \left\{A_1, A_2, \dots, A_n\right\}$, which contains n objects with d features. Data set A can be divided into $c(1 \le c \le n)$ clusters. Let $V_j(j = 1, 2, \dots, c)$ be the centre of the jth cluster and $U = \left(u_{ij}\right)_{n \times c}$ be the membership matrix of A_i to V_j. In the improved IFCM algorithm, the objection function can be stated as follows:

$$\begin{aligned} &\max J_{IFCM}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(u_{ij}\right)^m \bar{s}_{\omega}^2\left(A_i, V_j\right) \\ &s.t \sum_{j=1}^{c} u_{ij} = 1, \quad u_{ij} \in [0, 1], \quad 0 \le \sum_{i=1}^{n} u_{ij} \le n, \end{aligned} \tag{21}$$

where $\bar{s}_{\omega}\left(A_i, V_j\right)$ is the new aggregated intuitionistic fuzzy proximity measure proposed in Sect. 3.1, and the weight vector $\omega = \left(\omega_1, \omega_2, \dots, \omega_d\right)$ of features is obtained using the method presented in Sect. 3.2. The clustering results can be obtained through continuously updating u_{ij} and V_j by using Eqs. (14) and (22).

$$u_{ij} = \sum_{l=1}^{c} \left(\frac{\bar{s}_{\omega}\left(A_i, V_j\right)}{\bar{s}_{\omega}\left(A_i, V_l\right)}\right)^{\frac{1}{1-m}}. \tag{22}$$

We can observe many parameters in this clustering algorithm. According to [22], the weighting exponent (m) and Yager's intuitionistic fuzzy parameter (α) have important influences on the performance of the IFCM algorithm; suitable parameters can promote the success of the clustering algorithm. Therefore, parameter selection is crucial.

Amongst the optimization methods for selecting parameters (m and α), GA is used in this study because of its effectiveness in the global search of complex search spaces. As an evolution-based algorithm, GA represents each potential solution by using a chromosome-like data structure and then searches for an optimal solution via selection, crossover and mutation operators that imitate Darwinian natural evolution processes. The main mechanisms of the improved IFCM model optimized using the GA method are described as follows:

Step 1 (Initialization): Initial chromosome population ($popsize \times bit$) is randomly generated, where $popsize$ and bit are the numbers of chromosomes and parameters, respectively. Each chromosome represents a combination of two parameters (m and α), and it needs to be binary coded within a given range.

Step 2 (Evaluating fitness): The fitness of individuals in a population is calculated. We adopt Acc as the fitness function.

$$Acc = \frac{Count(|B \cap F|)}{N}. \tag{23}$$

where B refers to the correct clustering results, and F denotes the clustering results generated using the IFCM algorithm with the new aggregated intuitionistic fuzzy proximity measure. Accordingly, $Count(|B \cap F|)$ is the number of objects that are classified correctly using this IFCM algorithm. N represents the total amount of objects in the data set. Thus, the greater the fitness value is, the better the chromosome will be.

Step 3 (Selection): In order to ensure that the next generation of optimal chromosomes is better than the previous generation, the roulette wheel selection mechanism is used to retain several elite chromosomes with the highest fitness value. In this study, the number of retained elite chromosomes (*elist_n*) is set to 2.

Step 4 (Crosser and mutation): The chromosomes selected in the previous step are randomly matched to form parent pairs. In accordance with the single-point crossover principle, the middle segment between two randomly chosen break points is replaced. After crossover, some individuals are randomly selected to perform mutation operations with a certain probability. New individuals are generated by swapping 0 and 1 bit. The probabilities of crossover (*cross_rate*) and mutation (*mutate_rate*) need to be set in advance.

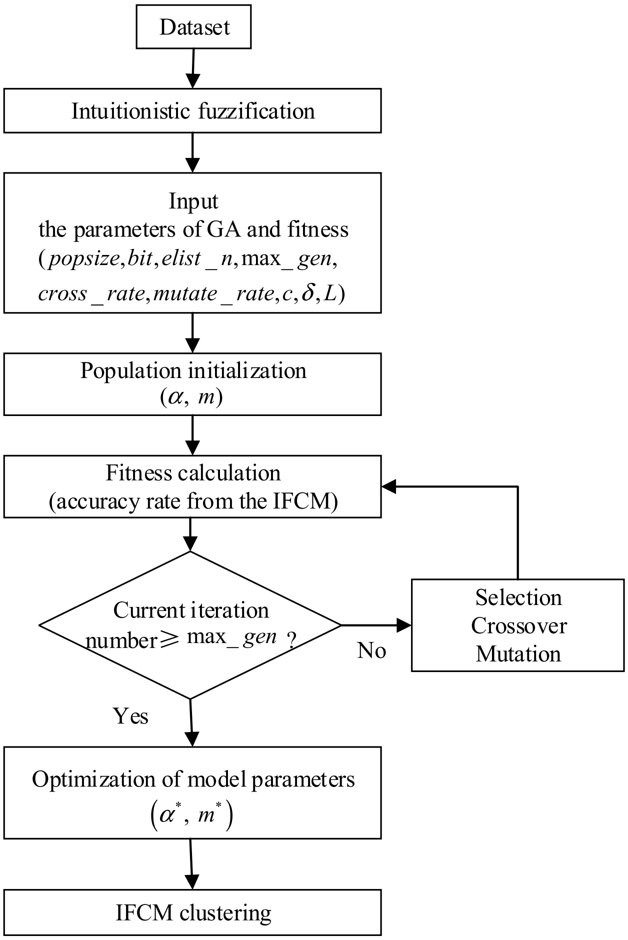

Dataset
Intuitionistic fuzzification
Input
the parameters of GA and fitness
$(popsize, bit, elist\_n, \max\_gen,$
$cross\_rate, mutate\_rate, c, \delta, L)$
Population initialization
(α, m)
Fitness calculation
(accuracy rate from the IFCM)
Current iteration
number$\geqslant \max_gen$?
No
Selection
Crossover
Mutation
Yes
Optimization of model parameters
(α^*, m^*)
IFCM clustering

Fig. 1 Flow chart of the proposed GA-IFCM algorithm

Step 5 (Next generation): After the crossover and mutation operations, the new generation population is formed. The evolution process needs to be repeated on the new population until the predefined stop criterion has been satisfied.

Step 6 (Stop criterion): If the number of generations reaches the given maximum genetic algebra ($\max\_gen$), then the optimal parameters and clustering results (U and V) of the IFCM algorithm with the optimal parameters are returned.

In short, the novel IFCM technique in this study adopts the aggregated weighted proximity measure, then GA is used to determine the optimal parameters in the clustering process. The flow chart of the GA-IFCM algorithm is shown in Fig. 1.

3.4 Analysis of computational complexity

In the IFCM algorithm, l, c, n and d are the numbers of iterations, clusters, objects and features, respectively. Each iteration in the IFCM method needs to calculate the distance between n objects and c cluster centres, which involves d dimensions. Hence, the time complexity of the IFCM algorithm is $o(ncdl)$. In GA, *popsize*, *bit* and $\max\_gen$ are the numbers of chromosomes, parameters to be optimized and genetic algebra, respectively. Each generation in GA needs to calculate the fitness value of all chromosome population ($popsize \times bit$). Thus, the time complexity of GA is $o(popsize \times bit \times \max\_gen)$. In the proposed GA-IFCM method, the IFCM algorithm is nested in each fitness calculation in the GA process. Therefore, the time complexity of the proposed method is $o(popsize \times bit \times \max\_gen \times ncdl)$.

4 Experimental analysis

In this section, we conduct three experiments on a synthetic data set and UCI machine learning data sets by using the proposed GA-IFCM algorithm and other comparison methods to demonstrate the effectiveness of the proposed clustering algorithm. Firstly, in order to illustrate the advantages of the proposed weighted proximity measure for IFSs, the traditional FCM algorithm and its variants which are based on different distance measures of crisp values, such as the kernel FCM (KFCM) algorithm [40] and Gustafson–Kessel (GK) algorithm [41], and six IFCM algorithms based on different similarity measures of IFSs are tested. Secondly, to clarify the role of GA in improving the performance of IFCM, we perform an experiment of optimising the parameters of IFCM with several heuristic optimization algorithms, including GA, PSO and ABC. Thirdly, we compare the proposed GA-IFCM algorithm with several state-of-the-art clustering algorithms, including partitioning, hierarchical,

Table 1 Summary of UCI data sets

Data set	No. of instances	No. of features	No. of classes	Feature characteristics
Iris	150	4	3	Real
Wine	178	13	3	Integer, real
Zoo	101	17	7	Categorical, integer
Breast Cancer Wisconsin	699	10	2	Integer
Dermatology	366	34	6	Categorical, integer

Table 2 Description of four cluster validity indexes

Validity index	Formula	Optical performance
PC [42]	$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\left(u_{ij}\right)^{2}$	max PC
SC [42]	$SC = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_{ij}\right)^{m}d_{ij}^{2}(x_i,v_j)}{N_j\sum_{k=1}^{c}d_{jk}^{2}(v_j,v_k)}$	min SC
S [42]	$S = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\left(u_{ij}\right)^{2}d_{ij}^{2}(x_i,v_j)}{N_j \min d_{jk}^{2}(v_j,v_k)}$	min S
DI [43]	$DI = \min_{j\in c}\left\{\min_{i\in c,i\neq j}\left\{\frac{\min_{x_i\in c_i,x_j\in c_j} d(x_i,x_j)}{\min_{x_i,x_j\in c} d(x_i,x_j)}\right\}\right\}$	max DI

density-based and model-based clustering, to prove its superiority. We conduct experiments on a personal computer with an Intel Core i5 2 GHz processor and 16 GB RAM.

Next, we will introduce the data sets, evaluation criteria and the experimental results.

4.1 Data sets

Experiments are conducted with a synthetic data set and five UCI data sets, namely, Iris, Wine, Zoo, Breast Cancer Wisconsin and Dermatology. The synthetic data set used in this study comes from a fuzzy clustering toolbox (Available: https://www.fmt.vein.hu/softcomp/fclusttoolbox/), which contains 2000 data points. A summary of the selected UCI machine learning data sets is given in Table 1.

The Iris and Wine data sets are the most commonly used data sets in clustering experiments. In the Zoo data set, the types of attributes are categorical and binary. The Breast Cancer Wisconsin data set comprises medical data which has a low number of classes. The Dermatology data set has a large number of features, and almost all of which are categorical variables. These data sets have different characteristics, thus, the experimental results can be comprehensively tested.

The IFCM algorithm is used to cluster IFSs. Hence, the real-valued UCI data sets should be transformed into IFSs through the intuitionistic fuzzification presented in Sect. 2.2.1.

4.2 Cluster evaluation indexes

In this study, we use accuracy, several cluster validation indexes and time consuming to evaluate the performance of each clustering algorithm.

Accuracy: It is a common index in machine learning, which is the ratio of the number of correctly classified objects to the total number of objects in a data set. The formula of accuracy is shown in Eq. (23).

Cluster validation indexes: Cluster validity indexes generally need to reflect intra-class compactness and interclass separation. Existing cluster validity indexes are often closely related to clustering tasks, and a universally perfect indicator is difficult to find. Thus, we combine several indexes to evaluate clustering results. Partition coefficient (PC) is a measure of the fuzziness of clustering. A large value of PC indicates a clear clustering. Partition index (SC) and separation index (S) are the ratios of intra-class variation to interclass separation, and Dunn index (DI) is the ratio of interclass separation to intra-class variation. Thus, small values of SC and S imply enhanced performance; the opposite is true for DI. The four cluster validity indexes we adopt are briefly described in Table 2.

The data set contains n objects and c clusters.

Time consuming: In data mining and machine learning, the time complexity is an important property of algorithms. High validity with a relatively short time is desirable.

4.3 Experiment 1: comparison of the proposed weighted proximity and other measurements

The objective of this experiment is to prove the advantages of the proposed weighted proximity measure by comparing with other measurement methods. The methods compared in this experiment are as follows: the FCM using the Euclidean distance of crisp values; the KFCM using the Gaussian kernel of crisp values; the GK using the Mahalanobis distance norm of crisp values; the conventional IFCM using the Euclidean distances of IFSs; the IFCM using the measurement functions shown in Sects. 2 and 3, including weighted Hamming distances $D_{\omega Ham}(A,B)$, Jiang's weighted distance $D_{\omega Jiang}(A,B)$, weighted Euclidean distances $D_{\omega E}(A,B)$, weighted tangent similarity $S_{\omega Tan}(A,B)$, the proposed new

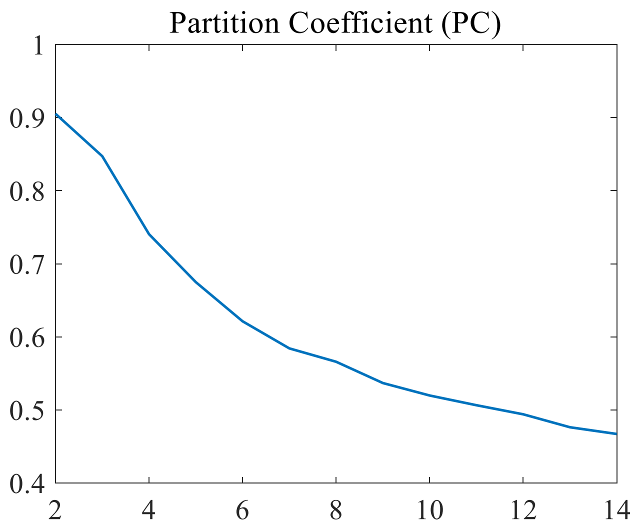


Fig. 2 Values of PC

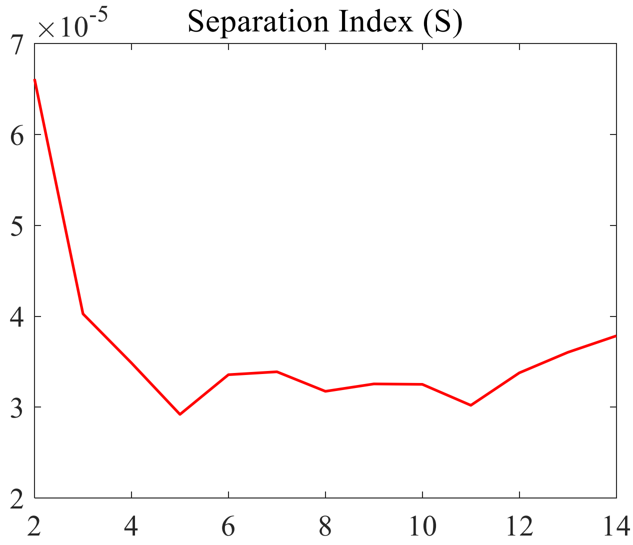


Fig. 3 Values of S

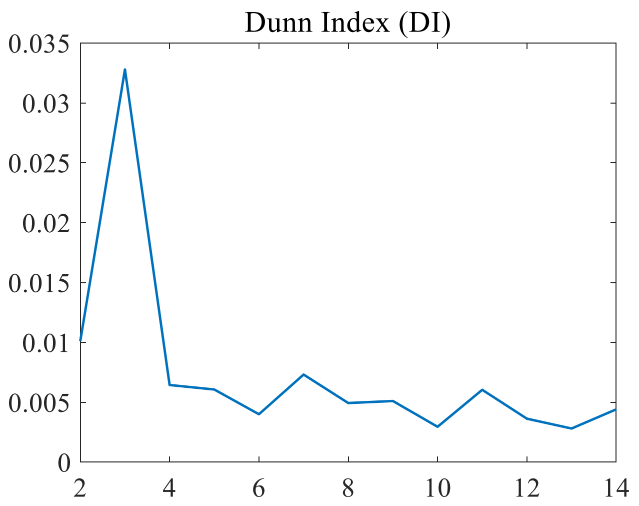


Fig. 4 Values of DI

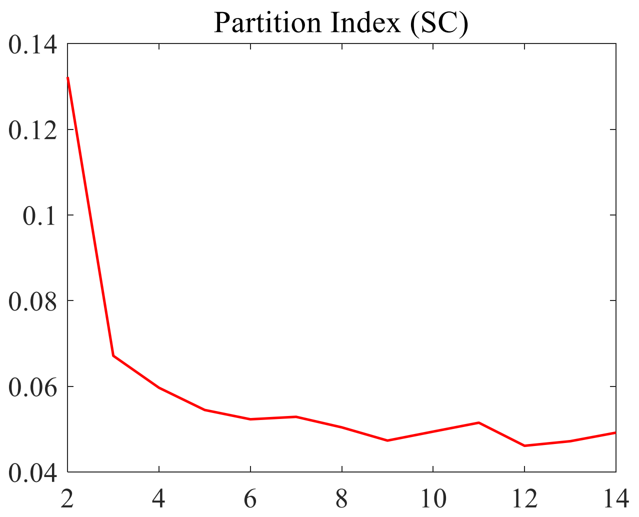


Fig. 5 Values of SC

weighted similarity measure $S_{\omega}(A, B)$ and the proposed weighted proximity measure $\overline{s}_{\omega}(A, B)$. Hereafter, they are expressed as FCM, KFCM, GK, IFCM_E, IFCM_wH, IFCM_wJiang, IFCM_wE, IFCM_wT, IFCM_wET and IFCM_wETC. In all experiments, we set $\delta = 0.0001$ and $L = 200$.

Before the clustering, the number of clusters c should be set in advance. But this task is difficult for the synthetic data set with an unknown data structure. In this case, we search for optimal c by using validity measures. Specifically, in the FCM algorithm, the value of c varies from 2 to 14, then the values of the validity indexes corresponding to each c are computed (Figs. 2, 3, 4, 5, 6, 7). The optimal value of c can be determined by analysing the changes in the values of all validity indexes.

From Fig. 2, we can observe that PC is monotonically decreasing with the increase in c. Obviously, the more the number of clusters divided is, the smaller the ambiguity of the clustering results is. Thus, the optimal value cannot be determined only on the basis of PC. Additional information is provided in Figs. 3 and 5. When $c < 3$, the values of S and SC decrease rapidly; when $c = 3$, they reach a local optimum (minimum); when $c > 3$, the values tend to be stable, and the extent of their decreases is small. Thus, from S and SC, $c = 3$ may be a good choice. Figure 4 shows that DI reaches its optimal (maximum) value when $c = 3$, which proves the superiority of the clustering results if c is regarded as 3. More intuitively, Figs. 6 and 7 show the original data points

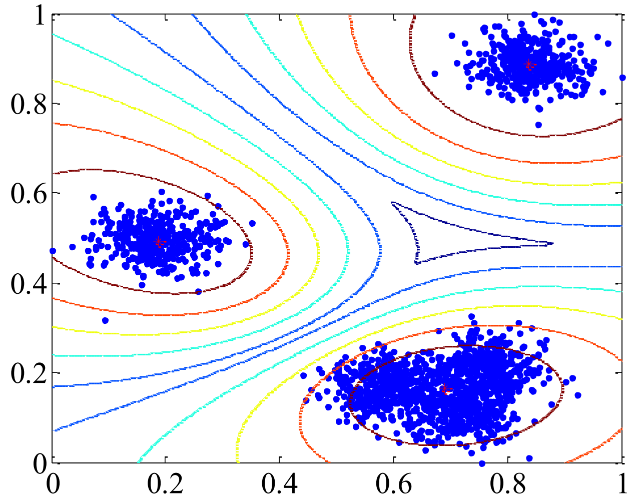


Fig. 6 Result of FCM with $c = 3$

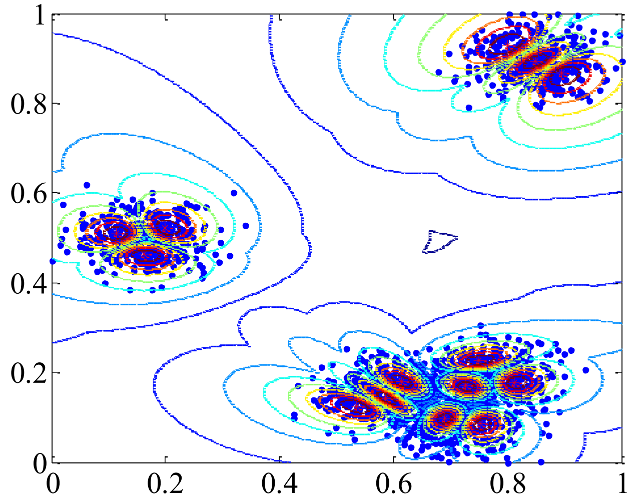


Fig. 7 Result of FCM with $c = 14$

and clustering results with $c = 3$ and $c = 14$, respectively. When $c = 3$, the clustering results are consistent with the distribution of the original data points; when $c = 14$, the 14 clusters are still divided into three categories in general. Therefore, in the case of the synthetic data set, we set the optimal number of clusters to 3.

Next, we perform an experiment with the methods listed above on the synthetic data set and several UCI data sets. Table 3, which is given in 'Appendix 1,' indicates the performance of each algorithm in accuracy, four cluster validation indexes and time consuming. The best results are shown in bold. On the basis of the results in Table 3, our analysis is as follows.

Firstly, as presented in Table 3, the clustering performance of FCM, KFCM and GK is poor. On most data sets, they have lower accuracy and higher SC and S compared with those of all IFCM methods. Although a high DI value is obtained, DI is unreliable in the case of overlapped clusters because of the redivision of the results with the hard division method. Therefore, accuracy, SC and S values demonstrate that clustering intuitionistic fuzzy data perform better than crisp data.

Secondly, except for the S of the Breast Cancer Wisconsin data set and the DI of Zoo and Dermatology data sets, almost all weighted IFCMs acquire better values of validation indexes than IFCM_E does. IFCM_E cannot consider the weight of features, resulting in information loss. The comparison results show that the feature weight in our proposed method should be taken into account.

Thirdly, compared with the IFCM algorithm using other similarity measures, IFCM_wETC obtains the highest accuracy and the best or second best PC, SC and S values for all data sets. Thus, the clustering results of IFCM_wETC are significantly better than those of other IFCM algorithms, which indicate that similarity and correlation should be aggregated as the proximity measure of two projects in clustering problems.

Lastly, in terms of time consuming, the FCM algorithm always spends the shortest time, but its accuracy is low. The IFCM algorithm consumes a longer time than FCM and KFCM do due to the complexity of the measurement method between two IFSs. Nevertheless, the difference is minimal and completely acceptable.

4.4 Experiment 2: comparison of GA and other parameter optimisation methods

After proving the superiority of the proposed weighted proximity measure, we analyse the performance of the parameter optimization method in improving the accuracy of IFCM. Amongst all heuristic optimization algorithms, GA, PSO and ABC are the most commonly used algorithms and have shown good capability in many parameter optimization problems. In this experiment, we employ them to optimize the parameters of IFCM_wETC and compare them in terms of accuracy and time consuming. Hereafter, they are expressed as GA_IFCM_wETC, PSO_IFCM_wETC and ABC_IFCM_wETC.

GA updates the search space by selecting, crossing and mutating chromosomes. The PSO algorithm finds the optimal solution by updating the position and velocity of particles. The ABC algorithm continuously searches and abandons food sources by worker, onlooker and scout bees. These heuristic algorithms require predefined parameters that have substantial effects on the results. With reference to previous studies [14], the parameter settings for this experiment are as follows. The crossover and mutation rates in GA are set to 0.7 and 0.1, respectively. The two learning rates for

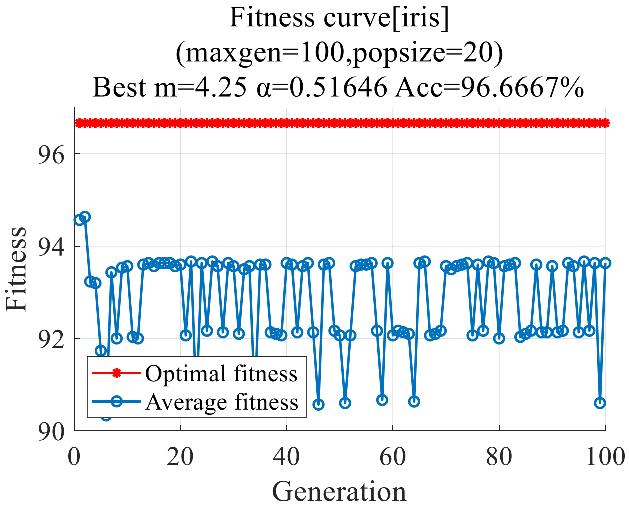


Fig. 8 Accuracy curve of GA_IFCM_wETC on Iris

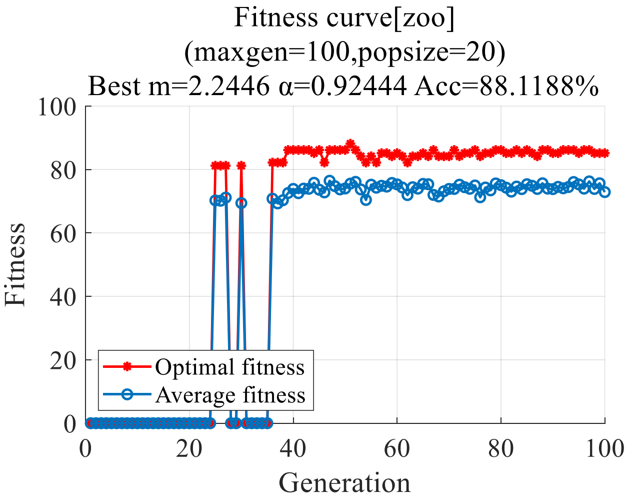


Fig. 10 Accuracy curve of GA_IFCM_wETC on Zoo

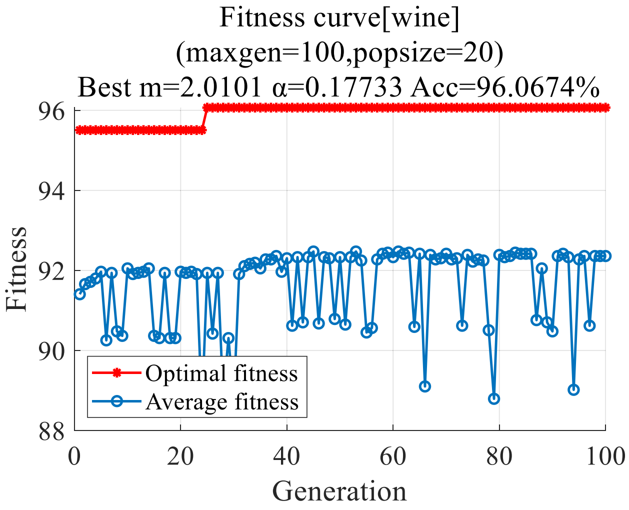


Fig. 9 Accuracy curve of GA_IFCM_wETC on Wine

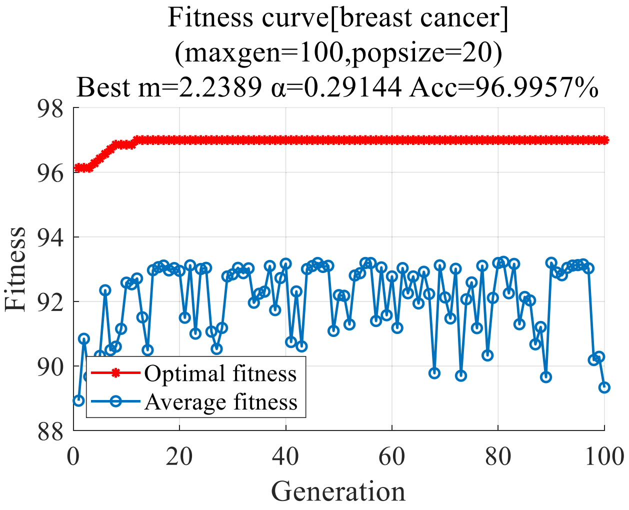


Fig. 11 Accuracy curve of GA_IFCM_wETC on Breast Cancer Wisconsin

particle updates in PSO are set to 2, and the limits of search and scout in ABC are set to 5. The number of populations in the three heuristic algorithms is set to 20. The maximum number of iterations in the three heuristic algorithms is set to 50 for the Dermatology data set, which has a large size and thus requires long processing time; meanwhile, the maximum number is set to 100 for all other data sets.

The comparison results of the three algorithms are shown in Table 4, which is given in ‘Appendix 1.’ Comparison of the results in Tables 3 and 4 shows that unlike the IFCM method based on parameters set manually, the three heuristic algorithms greatly improve the clustering results but take a longer time. On the one hand, in terms of accuracy shown in Table 4, GA_IFCM_wETC obtains high accuracy, especially for highly complex data sets, such as Zoo, Breast Cancer Wisconsin and Dermatology data sets. Compared with PSO and ABC algorithms, the mutation in GA can make it better to avoid falling into a local optimum. On the other hand, regarding time consuming shown in Table 4, with the same number of populations and iterations set, PSO_IFCM_wETC consumes the shortest calculation time and ABC_IFCM_wETC consumes the longest time in most data sets. The PSO algorithm has a simple search mechanism which only uses two parameters (the position and

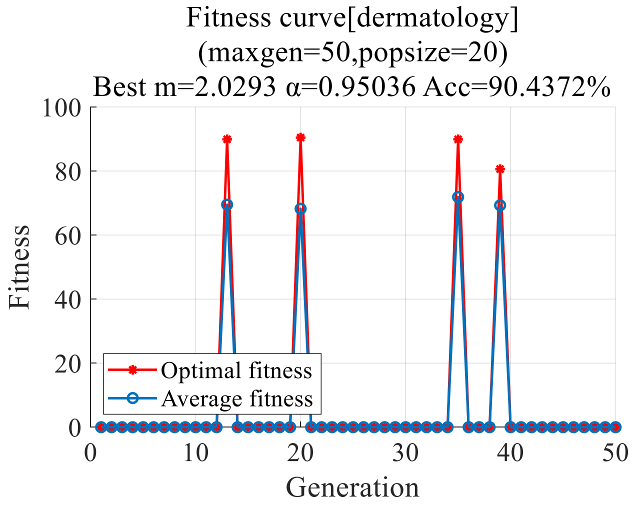


Fig. 12 Accuracy curve of GA_IFCM_wETC on Dermatology

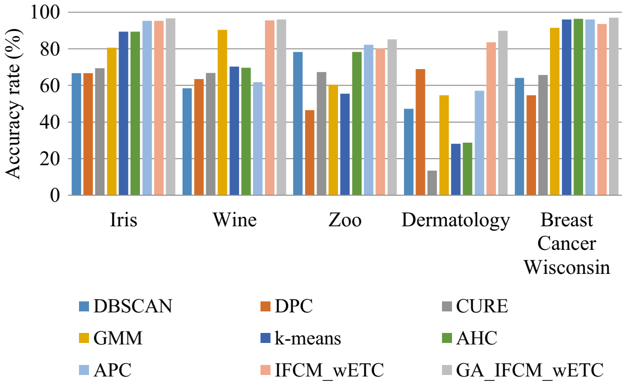


Fig. 13 Comparison of several well-known clustering algorithms on accuracy

velocity of particles), whereas the ABC algorithm needs to perform more calculations by three different types of bees (scout, employee and onlooker). GA has moderate time consuming, it spends less than 15 min on most data sets, which is acceptable. Overall, GA can achieve better results within a relatively short time, so using GA to determine optimal parameters can help traditional clustering algorithms obtain stable and accurate clustering results.

The accuracy curves of GA_IFCM_wETC on all data sets are shown in Figs. 8, 9, 10, 11 and 12.

4.5 Experiment 3: comparison of the proposed GA-IFCM and other clustering methods

In this experiment, the performance of the proposed GA-IFCM method using the proposed weighted proximity is compared with that of seven well-known clustering algorithms, namely, k-means [44], affinity propagation clustering (APC) [45], CURE [46], agglomerative hierarchical clustering (AHC) [47], DBSCAN [48], density peak clustering (DPC) [49] and Gaussian mixture model (GMM) [50], in clustering accuracy.

The comparison methods are introduced as follows. K-means algorithm is the most classic partitioning clustering algorithm, which groups data objects into k clusters in accordance with the nearest neighbour rule. The APC algorithm is one of the most competitive clustering algorithms proposed recently, which also belongs to partitioning clustering methods. It regards the similarity between pairs of data points as input and continuously updates the responsibility and availability of each point until the final exemplars are generated. It is stable on large-scale multi class data sets and is not easy to fall into local optima. The AHC algorithm is a bottom-up hierarchical clustering method. This method initially regards each data point as a class and then merges the data points with the highest similarity until the required number of clusters is reached. The calculation methods of similarity amongst clusters include ‘lingle-link’, ‘average-link’ and ‘complete-link.’ The ‘average-link’ is selected in this study. CURE is a hierarchical clustering algorithm based on representative points. The shrinking of representative points can reduce the effect of noise; hence, the algorithm can cluster data of arbitrary shapes. The DBSCAN algorithm is a typical density-based clustering algorithm, which divides continuous areas with high data density into one cluster. It determines the core point through two important parameters (the neighbourhood radius *Eps* and the minimum number of points *MinPts*), then links the core points with reachable density until all data points are divided. DPC is a novel clustering algorithm published in Science Journal in 2014, which mainly draws on the ideas of k-means, DBSCAN, mean shift and other algorithms. Its core idea is based on the assumption that the cluster centres are surrounded by neighbours with a low local density and have a large distance from other points with a high local density. It can efficiently detect noise on the basis of the density of the boundary area. Lastly, GMM is a popular model-based soft clustering algorithm, which can decompose the data set into several parts according to the Gaussian probability density function (normal distribution). Accordingly, it performs well on high-density data sets obeying normal distributions, but performs poorly on sparse clusters.

Figure 13 shows the clustering accuracy of our proposed method and the seven comparison methods on five UCI data sets. K-means and AP algorithms obtain good accuracy on Iris and Breast Cancer Wisconsin data sets because they can handle convex-shaped clusters well. But for arbitrary-shaped ones, the two algorithms are ineffective. The results of AHC and CURE are unstable because they are sensitive to parameters. Besides, CURE has no clear definition to remove noisy points. DBSCAN also faces the problem of parameter selection, and it cannot effectively handle high-dimensional data with considerable changes in density, such as Wine and Dermatology data sets. If a low-density cluster exists in a data set, the capability of DPC will be affected, for example, it acquires the lowest accuracy on the Zoo data set. GMM is sensitive to the initial value of the parameters. If the initialization is appropriate, good results can usually be achieved for data sets with normal distributions within the class, such as the Wine data set. However, different initialization values will lead to different clustering results, thereby greatly reducing the stability of GMM. The proposed IFCM clustering algorithm based on GA and the novel weighted proximity measure (denoted as GA_IFCM_wETC in Fig. 13) has the highest accuracy on all data sets. This finding proves that the proposed method considerably improves the clustering results. The improved GA-IFCM algorithm is more robust than other algorithms. Even without GA, the accuracy of the IFCM algorithm using the proposed weighted proximity measure (denoted as IFCM_wETC in Fig. 13) is also higher than that of most comparison methods. This result illustrates the superiority of the soft clustering algorithm based on the proposed novel weighted proximity.

From the research results of the above-mentioned three experiments, we can conclude that the proposed GA_IFCM_wETC may outperform other technologies in clustering problems. This superiority is mainly due to the advantage of the proposed weighted proximity measure in reducing information loss and the parameter optimisation capability of GA.

5 Conclusions

The FCM algorithm is an important soft clustering algorithm, which allows each object to be assigned to multiple clusters with varying membership degrees. IFSs can express valuable information; thus, the IFCM algorithm is developed to reduce the sensitivity of the FCM algorithm to noise and outliers. In the existing IFCM algorithm, the proximity degree between a pair of objects is always determined using the distance/similarity measures of IFSs, which only evaluates the closeness degree between the two projects and ignores the linear relationship between them. In addition, the parameters in the existing IFCM algorithm are always subjectively set by the user, which may lead to unreasonable clustering results. Therefore, we propose a novel weighted proximity measure in this paper to improve the IFCM clustering. The proposed measure combines the advantages of similarity and correlation measures to solve the first problem. To deal with the second problem, the improved IFCM algorithm applies GA to determine optimal parameters. Lastly, three experiments are conducted on a synthetic data set and UCI data sets. Experiments 1 and 2 prove that the proposed weighted proximity measure and GA perform better than other measurement methods and heuristic algorithms. The result of Experiment 3 demonstrates the superiority of the improved GA-IFCM algorithm to several latest clustering algorithms, including k-means, APC, AHC, CURE, DBSCAN, DPC and GMM.

In general, the clustering method proposed in this paper mainly has the following advantages. (1) A new weighted proximity measure that aggregates weighted similarity and correlation measures is proposed and applied to the IFCM algorithm. It can evaluate not only the closeness degree but also the linear relationship between two objects represented by IFSs. (2) The maximizing deviation method is used to calculate the weight of each feature in the new weighted proximity measure, which can reduce information loss for accurate clustering results. (3) GA is applied to the improved IFCM algorithm to determine the optimal parameters of the clustering algorithm effectively by thoroughly and optimally searching the solution space. In this way, the adverse effects of subjective parameter setting on the clustering results can be prevented.

However, the proposed algorithm has a limitation. Although GA can make clustering results feasible, it requires additional running time, especially in the case of processing large data sets. In the future, traditional heuristic algorithms need to be improved to optimize parameters or initialise cluster centres with minimal time. What's more, the application of the developed algorithm to other fields, such as data mining, image processing, and decision making, is also an interesting topic for future research.

Acknowledgements The authors are very grateful to the anonymous referees for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (no. 71871228).

Appendix 1

See Tables 3 and 4.

Table 3 Comparison of several soft clustering algorithms based on different measurement methods

Data set	Method	Parameters	Accuracy (%)	PC	SC	S	DI	Time (s)
Synthetic data set	FCM	m = 2,c = 3	—	0.9402	0.0367	2.2012e-05	**0.7336**	0.0354
	KFCM		—	0.3645	0.0047	2.8458e-06	0.0187	**0.1939**
	GK		—	0.9298	0.0372	2.2320e-05	0.5890	0.3745
	IFCM_E	α = 1,m = 2,c = 3	—	0.8600	0.0053	3.1670e-06	0.0868	0.1655
	IFCM_wH		—	0.8667	0.0049	2.9529e-06	0.0609	0.1600
	IFCM_wJiang		—	0.8645	0.0047	2.8411e-06	0.0868	0.1650
	IFCM_wE		—	0.8654	0.0051	3.0922e-06	0.0649	0.2365
	IFCM_wT		—	0.8652	0.0025	1.5204e-06	0.0700	0.1136
	IFCM_wET		—	0.8757	0.0034	2.0695e-06	0.0881	0.1297
	IFCM_wETC		—	**0.9342**	**0.0018**	**1.0544e-06**	0.0700	0.2207
Iris	FCM	m = 2,c = 3	89.3333	0.7834	0.1165	0.0012	0.1050	**0.0204**
	KFCM		71.3333	0.4101	0.0122	1.3696e-04	0.0363	0.0749
	GK		90	0.7277	0.0980	9.1845e-04	0.0413	0.4477
	IFCM_E	α = 1,m = 2,c = 3	89.3333	0.7425	0.0782	7.9328e-04	0.0347	0.1341
	IFCM_wH		95.3333	0.8336	0.0540	5.4520e-04	0.0347	0.0497
	IFCM_wJiang		95.3333	0.8336	0.0540	5.4520e-04	0.0347	0.0594
	IFCM_wE		96	0.8070	0.0659	6.6901e-04	0.1032	0.0818
	IFCM_wT		95.3333	0.8361	0.0283	2.8515e-04	0.0347	0.1129
	IFCM_wET		96	0.8171	0.0108	1.1081e-04	0.1073	0.1147
	IFCM_wETC		**96**	**0.8691**	**0.0065**	**6.5276e-05**	**0.1073**	0.1237
Wine	FCM	m = 2,c = 3	68.5393	**0.7909**	7.3291e-04	6.8364e-06	0.0117	0.0298
	KFCM		59.5506	0.5732	3.5055e-05	3.3227e-07	**0.0135**	**0.0819**
	GK		60.6742	0.3541	6.8649e-04	4.9870e-06	0.0019	0.0626
	IFCM_E	α = 1,m = 2,c = 3	91.0112	0.5462	1.9816e-06	1.5387e-08	0.0033	0.0687
	IFCM_wH		93.2584	0.5856	1.6115e-06	1.2501e-08	0.0035	0.7090
	IFCM_wJiang		93.2584	0.5856	1.6115e-06	1.2501e-08	0.0035	0.1027
	IFCM_wE		94.9438	0.5294	1.9362e-06	1.5043e-08	0.0039	0.0795
	IFCM_wT		93.8202	0.5898	8.3655e-07	6.4972e-09	0.0035	0.0567
	IFCM_wET		94.9438	0.5266	1.4107e-06	1.0919e-08	0.0034	0.0491
	IFCM_wETC		**95.5056**	0.6450	**7.4240e-07**	**5.8928e-09**	0.0039	0.0972
Zoo	FCM	m = 2,c = 7	54.4554	0.4521	0.1726	0.0025	**0.2182**	**0.0258**
	KFCM		53.4653	0.1506	0.0330	**4.4126e-05**	0.1925	0.1196
	GK		53.4653	0.4340	0.4318	0.0046	0.1459	0.9514
	IFCM_E	α = 1,m = 2,c = 7	69.3069	0.3601	0.0522	6.6821e-04	0.1213	0.0660
	IFCM_wH		73.2673	0.4825	0.0437	5.7273e-04	0.1195	0.3083
	IFCM_wJiang		75.2475	0.4776	0.0437	5.3791e-04	0.1195	0.6095
	IFCM_wE		66.3366	0.3738	0.0776	0.0010	01,187	0.0998
	IFCM_wT		81.1881	0.4527	0.0363	4.4106e-04	0.1187	0.3100
	IFCM_wET		79.2079	0.3983	0.0498	6.4063e-04	0.1187	0.2052
	IFCM_wETC		**82.1287**	**0.5285**	**0.0233**	3.2999e-04	0.1213	0.1617
Breast Cancer Wisconsin	FCM	m = 2,c = 2	95.7082	0.8364	0.0633	9.0572e-05	**0.1474**	0.0936
	KFCM		77.1102	0.5000	8.2307	0.0118	0.0387	0.7254
	GK		86.8383	0.7362	0.0733	1.0489e-04	0.0777	0.4019
	IFCM_E	α = 1,m = 2,c = 2	92.2747	0.8433	0.0513	7.3443e-05	0.1421	**0.0717**
	IFCM_wH		93.1330	0.8583	0.0398	5.6870e-05	0.1235	0.0828
	IFCM_wJiang		93.1330	0.8583	0.0398	5.6870e-05	0.1235	0.1209
	IFCM_wE		92.8469	0.8411	0.0498	7.1274e-05	0.1235	0.0790
	IFCM_wT		93.2761	0.8628	0.0013	**1.2505e-06**	0.0984	0.4276
	IFCM_wET		93.7053	0.8486	0.0015	2.1140e-06	0.0984	0.0847
	IFCM_wETC		**93.9914**	**0.8889**	**0.0011**	1.5651e-06	0.0984	0.3520

Table 3 (continued)

Data set	Method	Parameters	Accuracy (%)	PC	SC	S	DI	Time (s)
Dermatology	FCM	m = 2,c = 6	29.5082	**0.4659**	0.0171	7.4929e-05	0.1029	**0.0378**
	KFCM		30.0546	0.4626	**2.2253e-05**	**1.0423e-07**	**0.0727**	0.1433
	GK		54.6448	0.2890	12.4207	0.1174	0.0290	0.8808
	IFCM_E	α = 1,m = 2,c = 6	71.3115	0.2353	0.0107	3.4654e-05	0.0399	0.5731
	IFCM_wH		71.3115	0.2657	8.3945e-04	9.1248e-05	0.0254	0.7595
	IFCM_wJiang		71.3115	0.2657	9.8221e-04	9.1243e-05	0.2695	1.4600
	IFCM_wE		73.7705	0.3042	0.0082	2.6412e-05	0.0291	1.4941
	IFCM_wT		73.4973	0.3020	0.1531	4.8175e-04	0.3586	1.0652
	IFCM_wET		77.5956	0.2376	0.0059	1.9223e-05	0.0376	1.6910
	IFCM_wETC		**83.6006**	0.3515	0.0031	9.7574e-06	0.0309	2.1255

Table 4 Comparison of several heuristic optimisation algorithms

Data set	Method	Parameters	Accuracy (%)	Time (min)
Iris	GA_IFCM_wETC	α = 0.5165,m = 4.25,c = 3,maxgen = 100	**96.6667**	3.4842
	PSO_IFCM_wETC	α = 1,m = 2.4444,c = 3,maxgen = 100	**96.6667**	**3.2828**
	ABC_IFCM_wETC	α = 0.6619,m = 6.7222,c = 3,maxgen = 100	**96.6667**	4.1647
Wine	GA_IFCM_wETC	α = 0.1773,m = 2.0101,c = 3,maxgen = 100	**96.0674**	**7.5985**
	PSO_IFCM_wETC	α = 0.05,m = 1.1117,c = 3,maxgen = 100	**96.0674**	10.0748
	ABC_IFCM_wETC	α = 0.5988,m = 2.1189,c = 3,maxgen = 100	**96.0674**	21.7817
Zoo	GA_IFCM_wETC	α = 0.9244,m = 2.2446,c = 7,maxgen = 100	**88.1188**	14.1683
	PSO_IFCM_wETC	α = 1,m = 2,c = 7,maxgen = 100	86.1386	**6.2021**
	ABC_IFCM_wETC	α = 0.9605,m = 2.0751,c = 7,maxgen = 100	77.2277	16.3021
Breast Cancer Wisconsin	GA_IFCM_wETC	α = 0.2914,m = 2.2389,c = 2,maxgen = 100	**96.9957**	12.5514
	PSO_IFCM_wETC	α = 0.2914,m = 6.2,c = 2,maxgen = 100	**96.9957**	**6.9277**
	ABC_IFCM_wETC	α = 0.2995,m = 5.7079, c = 2,maxgen = 100	96.1373	13.1547
Dermatology	GA_IFCM_wETC	α = 0.9504,m = 2.0293,c = 6,maxgen = 50	**90.4372**	120.0517
	PSO_IFCM_wETC	α = 1,m = 2,c = 6,maxgen = 50	90.1104	73.3667
	ABC_IFCM_wETC	α = 0.9870,m = 3.6902,c = 6,maxgen = 50	74.8634	**23.2616**

References

1. Zadeh LA (1965) Fuzzy sets. Inf Comput 8:338–353
2. Atanassov KT (1986) Intuitionistic fuzzy sets. Fuzzy Sets Syst 20(1):87–96
3. Aruna Kumar SV, Harish BS (2018) A modified intuitionistic fuzzy clustering algorithm for medical image segmentation. J Intell Syst 27(4):593–607
4. Lingras P, West C (2004) Interval set clustering of web users with rough k-means. J Intell Inf Syst Integr Artif Intell Database Technol 23(1):5–16
5. Bai C, Zhang R, Qian L, Liu L, Wu Y (2018) An ordered clustering algorithm based on fuzzy c-means and PROMETHEE. Int J Mach Learn Cybern 10(6):1423–1436
6. Mitra S, Pedrycz W, Barman B (2010) Shadowed c-means: Integrating fuzzy and rough clustering. Pattern Recogn 43(4):1282–1291
7. Zhou J, Lai Z, Miao D, Gao C, Yue X (2020) Multigranulation rough-fuzzy clustering based on shadowed sets. Inf Sci 507:553–573
8. Yu H (2017) A framework of three-way cluster analysis. In: International joint conference on rough sets (IJCRS 2017). Springer, pp 300–312
9. Yu H, Zhang C, Wang G (2016) A tree-based incremental overlapping clustering method using the three-way decision theory. Knowl Based Syst 91:189–203
10. Ludwig SA (2015) MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability. Int J Mach Learn Cybern 6(6):923–934
11. Xu Z, Wu J (2010) Intuitionistic fuzzy C-means clustering algorithms. J Syst Eng Electron 21(4):580–590
12. Jain N, Kumar V (2016) IFCM based segmentation method for liver ultrasound images. J Med Syst 40(11):1–12

13. Son LH, Cuong BC, Lanzi PL, Thong NT (2012) A novel intuitionistic fuzzy clustering method for geo-demographic analysis. Expert Syst Appl 39(10):9848–9859
14. Kuo RJ, Lin TC, Zulvia FE, Tsai CY (2018) A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis. Appl Soft Comput 67:299–308
15. Fan X, Wang Y, Lei Y, Lu Y (2016) Long-term intuitionistic fuzzy time series forecasting model based on vector quantisation and curve similarity measure. IET Signal Proc 10(7):805–814
16. Keogh E, Ratanamahatana CA (2005) Exact indexing of dynamic time warping. Knowl Inf Syst 7(3):358–386
17. Karthikeyani Visalakshi N, Parvathavarthini S, Thangavel K (2014) An intuitionistic fuzzy approach to fuzzy clustering of numerical dataset. Adv Intell Syst Comput 246:79–87
18. Arora J, Khatter K, Tushir M (2019) Fuzzy c-means clustering strategies: a review of distance measures. Softw Eng 731:153–162
19. Milošević P, Petrović B, Jeremić V (2017) IFS-IBA similarity measure in machine learning algorithms. Expert Syst Appl 89:296–305
20. Lohani QMD, Solanki R, Muhuri PK (2018) Novel adaptive clustering algorithms based on a probabilistic similarity measure over atanassov intuitionistic fuzzy set. IEEE Trans Fuzzy Syst 26(6):3715–3729
21. Hwang CM, Yang MS, Hung WL (2018) New similarity measures of intuitionistic fuzzy sets based on the Jaccard index with its application to clustering. Int J Intell Syst 33(8):1672–1688
22. Chaira T (2011) A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. Appl Soft Comput 11(2):1711–1717
23. Gerstenkorn T, Mafiko J (1991) Correlation of intuitionistic fuzzy sets. Fuzzy Sets Syst 44:39–43
24. Hung W, Wu J (2002) Correlation of intuitionistic fuzzy sets by centroid method. Inf Sci 144(1):219–225
25. Hung WL (2001) Using statistical viewpoint in developing correlation of intuitionistic fuzzy sets. Int J Uncertainty Fuzziness Knowl Based Syst 9:509–516
26. Xu Z, Chen J, Wu J (2008) Clustering algorithm for intuitionistic fuzzy sets. Inf Sci 178(19):3775–3790
27. Liu B, Shen Y, Mu L, Chen X, Chen L (2016) A new correlation measure of the intuitionistic fuzzy sets. J Intell Fuzzy Syst 30(2):1019–1028
28. Wang F, Mao J (2018) Aggregation similarity measure based on intuitionistic fuzzy closeness degree and its application to clustering analysis. J Intell Fuzzy Syst 35(1):609–625
29. Nazari-Heris M, Mohammadi-Ivatloo B, Gharehpetian GB (2018) A comprehensive review of heuristic optimization algorithms for optimal combined heat and power dispatch from economic and environmental perspectives. Renew Sustain Energy Rev 21:2128–2143
30. Metawa N, Hassan MK, Elhoseny M (2017) Genetic algorithm based model for optimizing bank lending decisions. Expert Syst Appl 80:75–82
31. Chen S, Wang JQ, Zhang HY (2019) A hybrid PSO-SVM model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. Technol Forecast Soc Change 146:41–54
32. Huang CW, Lin KP, Wu MC, Hung KC, Liu GS, Jen CH (2014) Intuitionistic fuzzy c -means clustering algorithm with neighborhood attraction in segmenting medical image. Soft Comput 19(2):459–470
33. Lin KP (2014) A novel evolutionary kernel intuitionistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 22:1074–1087
34. Jiang Q, Jin X, Lee SJ, Yao S (2019) A new similarity/distance measure between intuitionistic fuzzy sets based on the transformed isosceles triangles and its applications to pattern recognition. Expert Syst Appl 116:439–453
35. Szmidt E, Kacprzyk J (2000) Distances between intuitionistic fuzzy sets. Fuzzy Sets Syst 114:505–518
36. Ye J (2011) Cosine similarity measures for intuitionistic fuzzy sets and their applications. Math Comput Model 53(1–2):91–97
37. Mondal K, Pramanik S (2015) Intuitionistic fuzzy similarity measure based on tangent function and its application to multi-attribute decision making. Glob J Adv Res 2(2):464–471
38. Bustince H, Burillo P (1996) Vague sets are intuitionistic fuzzy sets. Fuzzy Sets Syst 79(3):403–405
39. Wang Y (1997) Using the method of maximizing deviation to make decision for multiindices. J Syst Eng Electron 8(9):21–26
40. Zhang D, Chen S (2003) Clustering incomplete data using kernel-based fuzzy c-means algorithm. Neural Process Lett 18(3):155–162
41. Krishnapuram R, Kim J (1999) A note on the Gustafson–Kessel and adaptive fuzzy clustering algorithms. IEEE Trans Fuzzy Syst 7(4):453–461
42. Bensaid A, Hall LO, Bezdek JC, Clarke LP, Silbiger ML, Arrington JA, Murtagh R (1996) Validity-guided (re)clustering with applications to image segmentation. IEEE Trans Fuzzy Syst 4(2):112–123
43. Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern 3(3):32–57
44. Wang P, Shi H, Yang X, Mi J (2019) Three-way k-means: integrating k-means and three-way decision. Int J Mach Learn Cybern 10(10):2767–2777
45. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814):972–976
46. Guha S, Rastogi R, Shim K (2001) Cure: an efficient clustering algorithm for large databases. Inf Syst 26(1):35–58
47. Cheng D, Zhu Q, Huang J, Wu Q, Yang L (2018) A hierarchical clustering algorithm based on noise removal. Int J Mach Learn Cybern 10(7):1591–1602
48. Birant D, Kut A (2007) ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data Knowl Eng 60(1):208–221
49. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. Science 344(6191):1492–1496
50. Han X, Cui R, Lan Y, Kang Y, Deng J, Jia N (2019) A Gaussian mixture model based combined resampling algorithm for classification of imbalanced credit data sets. Int J Mach Learn Cybern 10(12):3687–3699