Improved Probabilistic Intuitionistic Fuzzy c-Means Clustering Algorithm: Improved PIFCM

Ayush K. Varshney
Department of Computer Science
South Asian University
New Delhi, India
varshneyayush90@gmail.com

Q. M. Danish Lohani
Department of Mathematics
South Asian University
New Delhi, India
danishlohani@cs.sau.ac.in

Pranab K. Muhuri
Department of Computer Science
South Asian University
New Delhi, India
pranabmuhuri@cs.sau.ac.in

Abstract— **Recently proposed Probabilistic Intuitionistic Fuzzy c-Means Algorithm (PIFCM) is a Probabilistic Euclidian distance measure (PEDM) based clustering technique, which incorporate computation of probabilistic intervals (P_{ij}, Q_{ij}) for each of the data point. PIFCM algorithm employs a random membership function $\frac{1}{|x|}$ and discards a data point if its membership value is uniformly distributed in the clusters. Fuzzy clustering always gets affected by the choice of the membership function. Accordingly, in PIFCM algorithm, membership function changes the properties of the data limiting its capabilities in giving consistent clustering results. Moreover, PIFCM algorithm incorporates computation of redundant matrices while finding P_{ij} and Q_{ij}. In this paper, we propose some novel changes in the existing PIFCM algorithm, and hence introduce our Improved PIFCM algorithm. The improved PIFCM algorithm considers the min-max normalization as membership function, and also removes the redundant matrix computation that was used to find the P_{ij} and Q_{ij} in the original PIFCM. Results over various UCI datasets validates the superiority of our improved PIFCM algorithm over FCM algorithm, IFCM algorithm and PIFCM algorithm.**

Keywords — Fuzzy clustering, AIFS based clustering, probabilistic interval, PEDM, PIFCM, IFCM.

I. Introduction

Clustering is an unsupervised technique for data analysis which partition the data into groups or subsets. Data elements in the same group share some common properties, and data elements in different groups mostly differs along all the properties. While clustering, maximum inter-cluster distance and minimum intra-cluster distance, is preferred. Clustering has been useful in various fields such as pattern recognition, data mining, information retrieval and so on [1], [2].

The concept of fuzzy sets (FSs), was introduced by Zadeh in 1965 [11]. FSs have been applied to the clustering domain by various researchers. Real-valued datasets does not account for the uncertainty present in the data. However, various variants of FSs, like type-2 Fuzzy sets [3], Atanassov Intuitionistic Fuzzy sets (AIFS) [4], interval type-2 fuzzy sets [5], vague sets [6], and many others [21], can depict the uncertainty in data to different extents.

Clustering is broadly classified as hard clustering or soft clustering. Hard clustering assigns data points to a single cluster i.e. each data can only belong to one cluster, while fuzzy or soft clustering assigns a data point to multiple clusters with a membership grade for a cluster. Membership grade determines the belongingness of data point to a particular cluster. Fuzzy clustering algorithms such as fuzzy c-means (FCM) [7], intuitionistic fuzzy c-means (IFCM) [8], [22] probabilistic intuitionistic fuzzy c-means (PIFCM) [9], intuitionistic fuzzy λ-cutting clustering algorithm [23] etc., has contributed significantly in the field of image analysis, data mining, and pattern recognition, etc.

FCM is an iterative algorithm based on the idea of k-means algorithm [10]. It re-computes the cluster centroids until the algorithm converges. In FCM, FSs are used to represent the data, and traditional Euclidian distance measure is used to compute the distance between two FSs. IFCM is an extension of FCM which uses AIFSs to represent the data [4]. It uses normalized Euclidian distance measure to compute distance between two AIFSs. Because AIFSs incorporate the uncertainty caused due to hesitancy, researchers have found AIFSs quite useful in clustering and many other machine learning problems [8], [9], [22]-[24].

The Recently developed FCM algorithms such as modified FCM [17], PIFCM [9], modified IFCM [18], modified IFCM incorporating hesitation degree [19], local information based improved IFCM [20], and many others improve either the objective function or the weight for distance measure or the constraints. Among them, the PIFCM is a recently proposed IFCM variant which also uses AIFSs to represent the data. The PIFCM algorithm uses probabilistic Euclidean distance measure (PEDM) [9] as proximity function (distance measure) to compute the distance between two data objects.

PEDM finds probabilistic weights for membership, non-membership and hesitance value from the dataset. Computed probabilistic weights are the mean of the mutual confidence interval between two data points. A mutual confidence interval can be seen as the agreement between two data objects. It is found through parameters P_{ij} and Q_{ij} proposed in PIFCM. Algorithms for P_{ij} and Q_{ij} compute confidence intervals for each of the data object and finds mutual confidence interval for membership and non-membership values, respectively, between each pair of data objects which is then used for distance computation. However, for any FCM based clustering algorithm, only the distance between a data point and the cluster centroids is computed. So, in PIFCM, redundant matrices are computed while computing P_{ij} and Q_{ij}, and hence increasing the cost and time of the algorithm.

Also, the membership function (MF) used in PIFCM was $\frac{1}{|x|}$, which is not a good choice, as it changes the properties present in the data. Mathematically, for $x \in [0,1]$, the membership value will be greater than 1, and for very large

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

valued x, membership value for x diminishes. How the MF $\frac{1}{|x|}$ changes the properties of data is shown in Fig. 1. Here, plot with blue line shows the original data, and plot with red line shows the transformation made by the MF $\frac{1}{|x|}$. Fig. 1(a) is the plot for $y = x$, whereas Fig. 1(b) is the plot for $y = sin(x)$. PIFCM algorithm also claims that data points should be excluded if membership values are uniformly distributed which is a bad practice. To compute mutual interval, we only need to find the probabilistic intervals between the data points and the cluster centroids for membership and non-membership components instead of the probabilistic intervals between each of the data points. We also need to use a different membership function which transforms the data into [0,1] but does not change the data properties.

Therefore, in this paper, we propose an improvement of the PIFCM algorithm, and accordingly term it as the 'improved PIFCM algorithm'. The improved PIFCM employs min-max normalization as membership function which transforms the data into [0,1]. It also ensures cost and space effective computation of P_{ij} and Q_{ij}. We have also included every datapoint in the computation, falsifying the claim of excluding the data points in case of uniformly distributed membership values in the PIFCM. Comparison among the FCM, IFCM, PIFCM and the proposed improved PIFCM algorithms have shown that the proposed algorithm outperforms its existing counterparts. In summary, the major contribution of the proposed algorithm are as follows:

1. A new cost and space effective technique to compute P_{ij} and Q_{ij} which in turn computes the PEDM effectively.
2. PIFCM algorithm claimed that the data points with uniformly distributed membership function values over more than one cluster should be excluded. However, this paper doesn't exclude any data points, and accordingly, formed clusters have shown good clustering accuracy.
3. Improved PIFCM algorithm employs min-max normalization function as the membership function which does not change the properties of the dataset, rather transform it to [0,1] without changing the data properties.

The rest of the paper is organized as follows: Section-II contains the pre-requisites required for the Improved PIFCM algorithm, Section-III contains the details of the proposed algorithm, Section-IV contains the experimental results and Section-V contains the summary and the future directions of the proposed work.

II. Pre-requisites

A. Fuzzy Sets (FSs): For an element x in the universe of discourse X, a FS A in regard to x can be defined as [11]:

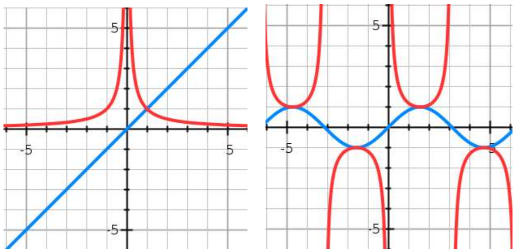
$$A = \{\langle x, \mu_A(x)\rangle | x \in X\}$$

Here, $\mu_A(x): X \to [0,1]$ is the membership function. The non-membership function $v_A(x)$ can be defined in terms of $\mu_A(x)$ as:

$$v_A(x) = 1 - \mu_A(x)$$

B. Atanassov Intuitionistic Fuzzy Sets (AIFSs): Let x be an element in the universe of discourse $X(\neq \emptyset)$. The AIFS A for the elements x can be defined as [4]:

$$A = \{\langle x, \mu_A(x), v_A(x)\rangle x \in X\} \qquad (1)$$



(a) (b)

Fig. 1 Change in data with $\frac{1}{|x|}$ membership function

Here, $\mu_A(x): X \to [0,1]$ is the membership function and $v_A(x): X \to [0,1]$ is the non-membership function which satisfies:

$$0 \le \mu_A(x) + v_A(x) \le 1$$

Hesitance degree, $\pi_A(x)$ of an element x is defined as:

$$\pi_A(x) = 1 - \mu_A(x) - v_A(x)$$

C. Distance Measure between AIFS: A function $d: AIFS(A) \times AIFS(B) \to [0,1]$ acts as a distance measure between two $AIFSs$ X and Y, if it satisfies the following properties [12]:

1) $0 \le d(X, Y) \le 1$
2) $d(X, Y) = 0$ only when $X = Y$
3) $d(X, Y) = d(Y, X)$
4) For AIFSs $X \subseteq Y \subseteq Z$, distance measure follows: $d(X, Z) \ge d(X, Y)$ and $d(X, Z) \ge d(Y, Z)$. (2)

D. Probabilistic Euclidian Distance Measure (PEDM): It is a Euclidean distance-based adaptive distance measure [9]. It computes probabilistic weights from the datasets. PEDM $\text{đ}_2(A_1, A_2)$ between two AIFSs A_1 and A_2 can be defined as follows:

$$\text{đ}_2(A_1, A_2) = \Big[\frac{1}{2n}\sum_{i=1}^{n}\Big(p_{12}\big(\mu_{A1}(x_i) - \mu_{A2}(x_i)\big)^2 + q_{12}\big(v_{A1}(x_i) - v_{A2}(x_i)\big)^2 + \rho(A_1, A_2)\big(\pi_{A1}(x_i) - \pi_{A2}(x_i)\big)^2\Big)\Big]^{\frac{1}{2}} \qquad (3)$$

Here, $p_{12} \in [p'_{min}, p''_{max}]$, $q_{12} \in [q'_{min}, q''_{max}]$ and $\rho(A_1, A_2)$ are the weights associated with the membership, the non-membership and the hesitance components, respectively. The intervals $[p'_{min}, p''_{max}]$ and $[q'_{min}, q''_{max}]$ are the confidence intervals and $\rho(A_1, A_2)$is the correlation between $A_1$and A_2. The intervals are computed as follows:

$$p'(A_{12}) = \max(p_{\min}(A_1), p_{\min}(A_2)),$$
$$p''(A_{12}) = \min(p_{\max}(A_1), p_{\max}(A_2))$$
$$q'(A_{12}) = \max(q_{\min}(A_1), q_{\min}(A_2)),$$
$$q''(A_{12}) = \min(q_{\max}(A_1), q_{\max}(A_2)).$$

E. AIFS generation

In the literature, there are quite a few techniques for generating AIFSs, e.g. Yager's generating function [13], [14], [25], and Meenakshi et al.'s IFS generating function [26]. In this paper, AIFS is generated for real valued data points using Yager's generating function. The generation technique is defined as follows:

$$\mu_A(x) = \zeta h(x), \quad \zeta \epsilon [0,1] \qquad (4)$$

$$v(x) = (1 - h(x)^{\alpha})^{\frac{1}{\alpha}} \qquad (5)$$

PIFCM algorithm choses $h(x)$ to be $\frac{1}{|x|}$ which causes loss of information. In this paper, we have chosen $h(x)$ as the min-max normalization i.e.

$$h(x) = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}$$

Also, for $\zeta = 1$, $\mu_A(x)$ from Eq. (5) may be as follows:

$$\mu_A(x) = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}} \tag{6}$$

and $\quad v_A(x) = \left(1 - \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}^{\alpha}\right)^{\frac{1}{\alpha}}$

Hence, using eq. (1) AIFS A is written as:

$$A = \left\{ \left\langle x, \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}, \left(1 - \frac{x_i - x_{imin}}{x_{imax} - x_{imin}}^{\alpha}\right)^{\frac{1}{\alpha}} \right\rangle x \in X \right\} \tag{7}$$

F. Intuitionistic Fuzzy c-Means Algorithm (IFCM)

Euclidian Distance measure [15] acts as the proximity function in IFCM [8]. In IFCM algorithm, AIFSs are used to represent the real valued data points. IFCM clusters p data points, where each data point is in n-th dimension, into c clusters. The objective function for IFCM is given below:

$$\begin{gathered} \min J_m = \textstyle\sum_{i=1}^{p} \sum_{j=1}^{c} u_{ij}^{m} \acute{d}_{ij}^{2} \\ \text{s.t.} \quad \textstyle\sum_{j=1}^{c} u_{ij} = 1, 1 \le j \le c \\ u_{ij} \ge 0, 1 \le i \le p, 1 \le j \le c \\ \sum_{i=1}^{p} u_{ij} > 0, 1 \le j \le c \end{gathered} \tag{8}$$

Here, u_{ij} acts as a partition matrix which contains membership of i-th data point into j-th cluster, $\acute{d}_{ij}^{2}$ is the Euclidian distance measure which is used to compute the distance between i-th data point and j-th cluster centroid, and $m \in [1, \infty]$is the fuzzifier constant that controls how fuzzy the cluster will be.

G. Probabilistic Intuitionistic Fuzzy c-Means Algorithm (PIFCM)

PIFCM is an extension of IFCM algorithm, it uses PEDM as the proximity function [9]. Each of the n-dimension of the data point is represent by using an AIFS. PIFCM algorithm also clusters p data points into c clusters. The objective function for PIFCM algorithm is given below:

$$\begin{gathered} \min J_m = \textstyle\sum_{i=1}^{p} \sum_{j=1}^{c} u_{ij}^{m} \text{đ}_{ij}^{2} \\ \text{s.t.} \quad \textstyle\sum_{j=1}^{c} u_{ij} = 1, 1 \le j \le c \\ u_{ij} \ge 0, 1 \le i \le p, 1 \le j \le c \\ \sum_{i=1}^{p} u_{ij} > 0, 1 \le j \le c \end{gathered} \tag{9}$$

Here, đ_{ij}^{2} is the PEDM which also computes the distance between i-th data point and j-th cluster centroid, u_{ij} is the partition matrix which has entries of belongingness of i-th data point into j-th cluster and m is the fuzzifier constant.

H. Algorithm for finding P_{ij} [9]

Pij is a parameter used as a weight for membership component in the PEDM. It is computed as follows:

Step 1: Assign membership for each of the n attributes of p data objects. Find the set $C = \{\mu'(x_1), \mu'(x_2), \dots, \mu'(x_n)\}$ which contains the set of minimum membership value for each attribute.

Step 2: Find the sum of the set C, S=$\sum_{i=1}^{n} \mu'(x_i)$.

Step 3: If $S \neq 0$, then for each attribute $p(x_i) = \frac{\mu'(x_i)}{S}$; otherwise, $p(x_i) = 1$.

Step 4: Compute the minimum probability for each object A_j.

$$p_{min}(A_j) = \sum_{i=1}^{n} p(x_i)\mu_{A_j}(x_i)$$

Step 5: Compute maximum probability for A_j as follows:

$$p_{max}(A_j) = p_{min}(A_j) + \sum_{i=1}^{n} p(x_i)\pi_{A_j}(x_i)$$

Step 6: Derive the mutual confidence interval between each of the data objects (for A_i and A_j) as follows:

$$\begin{aligned} p_{ij} &= p(A_i, A_j) \\ &= \begin{cases} [p'(A_{ij}), p''(A_{ij})] \text{ if } p'(A_{ij}) \le p''(A_{ij}) \\ [p''(A_{ij}), p'(A_{ij})], \quad \text{otherwise} \end{cases} \end{aligned}$$

I. Algorithm for finding Q_{ij} [9]

Qij is a parameter used as a weight for non-membership component in the PEDM. It is computed as follows:

Step 1: Compute the non-membership for each of the n attributes of p data objects using Eq. (7). Find the set $C = \{v'(x_1), v'(x_2), \dots, v'(x_n)\}$ which contains the set of minimum membership value for each attribute.

Step 2: Find the sum of the set C, $S' = \sum_{i=1}^{n} v'(x_i)$.

Step 3: if $S \neq 0$, then for each attribute $q(x_i) = \frac{v'(x_i)}{S'}$; otherwise, $q(x_i) = 1$.

Step 4: Compute the minimum probability for each object A_j.

$$q_{min}(A_j) = \sum_{i=1}^{n} q(x_i)\mu_{A_j}(x_i)$$

Step 5: Compute maximum probability for A_j as:

$$q_{max}(A_j) = q_{min}(A_j) + \sum_{i=1}^{n} q(x_i)\pi_{A_j}(x_i)$$

Step 6: Derive mutual confidence interval between each of the data objects (for A_i and A_j) as:

$$\begin{aligned} q &= q(A_i, A_j) \\ &= \begin{cases} [q'(A_{ij}), q''(A_{ij})] \quad \text{if } q'(A_{ij}) \le q''(A_{ij}) \\ [q''(A_{ij}), q'(A_{ij})], \quad \text{otherwise} \end{cases} \end{aligned}$$

III. Proposed Work

In this section, we have explained our proposed Improved Probabilistic Intuitionistic Fuzzy c-Means (Improved PIFCM) Algorithm. This section also contains the algorithms to find computational and space effective probabilistic intervals between the data points and the cluster centroids. The algorithms used to find the probabilistic intervals for the data points are given in the Section III-A, whereas in the Section

III.B, we provide the flowchart and the pseudocode of the Improved PIFCM algorithm.

A. *Probabilistic intervals of the data points*

To compute the P_{ij} and Q_{ij}, defined in the Sections II.H and II.I, using the membership function defined in Eq. (6), each of the attribute will have 0 (in case of minimum value) and 1 (in case of maximum value) membership value in any one of the p data points. So, the $p(x_i)$ for P_{ij} and $q(x_i)$ for Q_{ij} component in the PIFCM algorithm will always be 1 for each attribute, and hence will have a fixed value. Also, we do not need to compute the confidence interval between two data objects; rather between a data object and the cluster centroids. Hence, Step 6 in both of the algorithms is redundant.

Algorithm 1 computes the confidence interval $[p_{min}, p_{max}]$ for the membership values. The mutual confidence between a data point and the cluster centroids can be computed while computing the PEDM. The mean of the mutual interval is taken as the weight for the membership component.

```
Algorithm 1: Algorithm for finding the [p_min, p_max]
1: Find the minimum probability value for each object $A_j (j = 1,2,3 \ldots, p), p_{min}(A_j)$ as follows:
       $p_{min}(A_j) = \sum_{i=1}^{n} \mu_{A_j}(x_i)$.
2: Compute the maximum probability value for each object $A_j$, $p_{max}(A_j)$ as follows:
       $p_{max}(A_j) = p_{min}(A_j) + \sum_{i=1}^{n} \pi_{Aj}(x_i)$,
   where $\pi_{Aj}(x_i) = 1 - \left(\mu_{Aj}(x_i) + v_{Aj}(x_i)\right)$.
```

Similarly, we can compute the confidence interval $[q_{min}, q_{max}]$ for the non-membership value using the *Algorithm 2*.

```
Algorithm 2: Algorithm for finding the [q_min, q_max]
1: Compute the minimum probability value for each object $A_j (j = 1,2,3 \ldots, p), q_{min}(A_j)$ as follows:
       $q_{min}(A_j) = \sum_{i=1}^{n} v_{A_j}(x_i)$.
2: Calculate the maximum probability value for each object $A_j, q_{max}(A_j)$ as follows:
       $q_{max}(A_j) = q_{min}(A_j) + \sum_{i=1}^{n} \pi_{Aj}(x_i)$,
   where $\pi_{Aj}(x_i) = 1 - \left(\mu_{Aj}(x_i) + v_{Aj}(x_i)\right)$.
```

Improvement in computational time and space:

Probability measure defined in the PIFCM algorithm has a running time of $O(p^2 n)$ while the running time for the Improved PIFCM proposed here is $O(pnc)$. The space requirement for PIFCM is $O(p^2)$ while in our proposed method, it is $O(pc)$.

B. *Improved Probabilistic Intuitionistic Fuzzy c-Means Algorithm (Improved PIFCM):*

Here, we introduce our proposed Improved Probabilistic Intuitionistic Fuzzy c-Means (Improved PIFCM) Algorithm. Improved PIFCM uses the confidence intervals $[p_{min}, p_{max}]$ and $[q_{min}, q_{max}]$ proposed in the Section III.A to find the mutual confidence for weights in the PEDM. It uses membership function defined in Eq. (6) instead of $\frac{1}{|x|}$ of the PIFCM. Fig. 1 gives the flowchart for the proposed Improved PIFCM algorithm and *Algorithm 3* gives its pseudocode.

Along with the PIFCM algorithm, the proposed Improved PIFCM algorithm is also an iterative algorithm which computes new seeds with each iteration until the algorithm converges. Here, $\bar{d}_2^2$ is the PEDM defined in Eq. (3), u_{ij} is the partition matrix, $S(x)$ is the set of cluster centroid at x iteration, $avgop$ is the average operator used to compute the next set of cluster centroids.

Improved PIFCM algorithm is an extension of PIFCM which finds better clusters efficiently. First, the Improved PIFCM algorithm randomly initializes cluster centroids, then computes the partition matrix over these cluster centroids. It repeatedly computes the cluster centroids and the partition matrix until the algorithm converges.

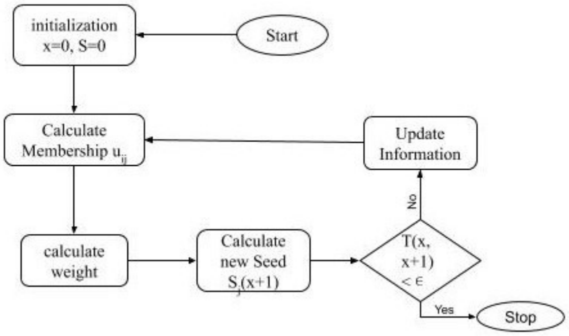


Fig. 1 Flowchart of the Improved PIFCM

```
Algorithm 3: Improved PIFCM Algorithm
1: Initialize $m, \in, c, A, x = 0$and choose initial seeds $S(0)$
2: Compute partition matrix, $M(r) = \left(u_{ij}(r)\right)_{p \times c}$ such that,
   a) If $\exists j \wedge i$ for which $\bar{d}_2^2\left(A_i, S_j(r)\right) = 0$, then assign $u_{ij} = 1$ and $u_{ij} = 0 \forall j \neq k$.
   b) Otherwise,
      $u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\bar{d}_2^2(A_i, S_j)}{\bar{d}_2^2(A_i, S_k)}\right)^{\frac{2}{m-1}}}$
3: Compute the next seeds $S(x+1)$,
       $S(x+1) = \{S_1(x+1), S_2(x+2), \ldots, S_c(x+1)\}$
       $S_i(x+1) = avgop\left(A, w_j(x+1)\right), 1 \le j \le c$,
   where $avgop$ is the average operator of PIFCM.
   Also,   $w_j = \left\{\frac{u_{ij}^m}{\sum_{i=1}^{p} u_{ij}^m}, 1 \le i \le p\right\}$.
4: Check if $T(x, x+1) = \sum_{k=1}^{c} \frac{\bar{d}_2(S_k(r), S_{k+1}(r))}{c} < \epsilon$. If it holds, jump to Step 5; otherwise, increment $x$ and jump to Step 3.
5: END.
```

IV. Experimental Results

In this section, we have explained the results of our Improved PIFCM algorithm over various UCI datasets. We have compared the proposed improved PIFCM with FCM, IFCM, PIFCM algorithms. We have compared the time taken by the PIFCM and the Improved PIFCM algorithms to show that proposed algorithm is computationally feasible and advantageous. Table I gives the details of the datasets used for the comparison.

A. *Clustering Accuracy over various UCI datasets:*

Clustering accuracy: Clustering accuracy is one of the most used criteria to compare the results of a clustering algorithm. We have used clustering accuracy to compare the performance of FCM, IFCM, PIFCM and proposed Improved PIFCM. Mathematically, clustering accuracy can be defined as follows:

$$\text{Clustering accuracy} = \frac{Numberofcorrectlyclassifiedsamples}{totalnumberofsamples}$$

TABLE I. USED UCI DATASETS

Dataset	No of instances	No of features	No of classes
Balance scale	625	4	3
Breast Cancer	569	30	2
Car Evaluation	1728	6	4
Dermatology	366	34	6
Ecoli	336	7	8
Glass	214	9	6
Image Segmentation	2310	19	7
Iris	150	4	3
Ionosphere	351	33	2
Seeds	210	7	3
Wine	178	13	3
Zoo	101	17	7

TABLE II. CLUSTERING ACCURACY OVER VARIOUS UCI DATASETS

Datasets	**FCM (%)**	**IFCM (%)**	**PIFCM (%)**	**Improved PIFCM (%)**
IRIS	90.67	92.67	94.00	**94.67**
	m=2.1	m=3.6, α=0.55	m=1.1, α=0.50	m=3.7, α=0.40
ZOO	83.168	92.08	92.08	**92.08**
	m=1.8	m=1.5, α=0.40	m=1.6, α=0.05	m=1.6, α=0.60
WINE	93.25	93.82	**56.18**	**96.07**
	m=3.9	m=2.7, α=0.85	m=3.9, α=0.65	m=1.3, α=0.05
BREAST CANCER	92.09	94.20	**62.74**	**94.38**
	m=1.2	m=1.8, α=0.35	m=1.1, α=0.05	m=2.8, α=0.35
BALANCE SCALE	71.2	77.76	68.16	**79.2**
	m=1.9	m=1.7, α=0.95	m=1.1, α=0.05	m=2.0, α=0.75
SEEDS	90.47	**91.43**	**63.81**	90.48
	m=2.7	m=1.1, α=0.65	m=4.0, α=0.50	m=3.6, α=0.55
IMAGE SEGMEN-TATION	64.81	69.28	**28.76**	**69.67**
	m=2.8	m=2.3, α=0.55	m=2.8, α=0.05	m=2.1, α=0.55
Car Evaluation	76.04	81.08	70.03	**81.14**
	m=1.2	m=1.5, α=0.5	m=1.3, α=0.55	m=1.1, α=0.80
Dermatology Dataset	89.34	92.90	92.07	**92.90**
	m=2.8	m=1.4, α=0.05	m=1.3, α=0.05	m=1.3, α=0.05
Ionosphere	70.94	71.23	64.39	**71.23**
	m=1.1	m=2.3, α=0.65	m=1.7, α=0.20	m=1.2, α=0.7
Ecoli	79.46	84.82	73.51	**84.82**
	m=3.6	m=1.6, α=0.30	m=2.7, α=0.05	m=1.5, α=0.50

Table II shows the clustering accuracies of the algorithms over various UCI datasets [16]. From the Table II it can be seen that, except the SEEDS dataset, clustering result of our proposed improved PIFCM algorithm outperforms all the existing counterparts in terms of the clustering accuracy. For the SEEDS dataset, IFCM algorithm gives better clustering accuracy than other three algorithms.

- *Why PIFCM algorithm performs poorly in most of the datasets present in the Table II?*

The poor performance of the PIFCM algorithm over Wine, Breast Cancer, Seeds, Image Segmentation and other datasets is because of the membership function it used. PIFCM uses the membership function of $\frac{1}{|x|}$. In the above mentioned datasets, there exists a number of difficulties with this membership function, such as:

- Many attributes of the above mentioned datasets have the domain of $[0,1]$. Since, the values of the attributes are less than 1, the membership value will be greater than 1, which is a contradiction. Also, this results in negative non-membership value, which leads to negative weights for PEDM and complex distance between the data points. Accordingly, it affects the clustering accuracy.
- Many attributes in the datasets have the domain consisting of positive and negative values. The membership values of negative real-valued data is mixed with the positive real-valued data, and hence causes low clustering accuracy.

Table III comparatively provides the running time of the PIFCM algorithm and the proposed improved PIFCM algorithm. From the Table III, we can see that running time of proposed PIFCM algorithm is significantly less than the running time of the PIFCM algorithm. It supports that proposed probabilistic interval in the Section III.A is computationally feasible.

TABLE III. COMPARISON OF RUNNING TIMES OF THE PIFCM ALGORITHM AND THE PROPOSED IMPROVED PIFCM ALGORITHM

Datasets	**PIFCM (*Seconds*)**	**Improved PIFCM (*Seconds*)**
IRIS	0.09	**0.05**
ZOO	0.17	**0.15**
WINE	4.84	**0.07**
BREAST CANCER	1.54	**0.55**
BALANCE SCALE	1.42	**0.35**
SEEDS	0.31	**0.18**
IMAGE	73.82	**3.35**
Car Evaluation	27.44	**1.04**
Dermatology Dataset	1.84	**0.96**
Ionosphere	1.16	**0.14**
Ecoli	1.41	**0.78**

Consider an example for Image Segmentation dataset, where, no of data points = 2310. Here, PIFCM algorithm finds redundant matrices while computing P_{ij} and Q_{ij}. Size of the redundant matrix = 2310×2310. Also, PIFCM algorithm computes two matrices, each of size 2310×2310 in each iteration. Then it computes the confidence intervals between the data points and the cluster centroids of size 2310×7. On the other hand, proposed Improved PIFCM algorithm computes mutual confidence intervals between the data points and the cluster centroids of size 2310×7. Therefore, proposed

Improved PIFCM algorithm computes 10,672,200 $(2 \times 2310 \times 2310)$ less mutual confidence intervals and hence, computationally is more efficient than the existing PIFCM algorithm.

V. Summary

In this paper, we have proposed the Improved PIFCM algorithm which improves the recently proposed PIFCM algorithm over various fronts. Here, we have shown the drawbacks of the PIFCM algorithm and have proposed techniques which are computationally effective. Efficiently computed weights have also led to the efficient computation of the PEDM. Proposed algorithm highlights the importance of choosing a good membership function and creates better clustering results. We have chosen min-max normalization as the membership function; a better membership function can also be incorporated. Improved PIFCM algorithm along with PIFCM algorithm is highly dependent on cluster centroids initialization. Improved PIFCM algorithm fails in the case of noisy data. A normally distributed-PIFCM algorithm may create good clustering results. So, initialized and normally distributed Improved PIFCM algorithm can be a good direction for future work.

References

[1] S. B. Everitt, S. Landau, and M. Leese. "Cluster analysis arnold." *A member of the Hodder Headline Group, London*, pp. 429-438, 2001.

[2] W. Pedrycz. “Knowledge-based clustering: from data to information granules,” *John Wiley & Sons*, 2005.

[3] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning—I,” *Information Sciences.*, vol. 8, no. 3, pp. 199–249, 1975.

[4] K. T. Atanassov. "Intuitionistic fuzzy sets," In *Intuitionistic fuzzy sets*, pp. 1-137, Physica, Heidelberg, 1999.

[5] J. M. Mendel, R. I. John, and F. Liu. "Interval type-2 fuzzy logic systems made simple." *IEEE transactions on fuzzy systems*, vol. 14, no. 6, pp. 808-821, 2006.

[6] W-L. Gau, and D. J. Buehrer. "Vague sets," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 2, pp. 610-614, 1993.

[7] J. C. Bezdek, R. Ehrlich, and W. Full. "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[8] Z. Xu, and J. Wu. "Intuitionistic fuzzy c-means clustering algorithms," *Journal of Systems Engineering and Electronics*, vol. 21, no. 4, pp. 580-590, 2010.

[9] Q. M. D. Lohani, R. Solanki, and P. K. Muhuri. "Novel adaptive clustering algorithms based on a probabilistic similarity measure over atanassov intuitionistic fuzzy set," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 6, pp. 3715-3729, 2018.

[10] J. A. Hartigan, and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[11] L. A. Zadeh, "Fuzzy sets," *Information and control* vol. 8, no. 3, pp. 338-353, 1965.

[12] W. Wang, and X. Xin. "Distance measure between intuitionistic fuzzy sets," *Pattern Recognition Letters*, vol. 26, no. 13, pp. 2063-2069, 2005.

[13] R. R. Yager, "On the measure of fuzziness and negation part I: membership in the unit interval," pp. 221-229, 1979.

[14] R. R. Yager. "On the measure of fuzziness and negation. II. Lattices," *Information and control*, vol. 44, no. 3, pp. 236-260, 1980.

[15] E. Szmidt, and J. Kacprzyk. "Distances between intuitionistic fuzzy sets," *Fuzzy sets and systems*, vol. 114, no. 3, pp. 505-518, 2000.

[16] A. Asuncion, and D. Newman. "UCI machine learning repository," 2007.

[17] D. Chakraborty, and S. Das. "Modified fuzzy c-mean for custom-sized clusters," *Sādhanā*, vol. 44, no. 8, pp. 182, 2019.

[18] D. Kumar, H. Verma, A. Mehra, and R. K. Agrawal. "A modified intuitionistic fuzzy c-means clustering approach to segment human brain MRI image," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 12663-12687, 2019.

[19] H. Verma, A. Gupta, and D. Kumar. "A modified intuitionistic fuzzy c-means algorithm incorporating hesitation degree," *Pattern Recognition Letters*, vol. 122, pp. 45-52, 2019.

[20] H. Verma, R. K. Agrawal, and A. Sharan. "An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation." Applied Soft Computing 46 (2016): 543-557.

[21] H. Bustince, E. Barrenechea, M. Pagola, J. Fernandez, Z. Xu, B. Bedregal, J. Montero, H. Hagras, Francisco Herrera, and Bernard De Baets. "A historical account of types of fuzzy sets and their relationships." *IEEE Transactions on Fuzzy Systems* 24, no. 1 (2015): 179-194.

[22] Q. M. D. Lohani, R. Solanki, and P. K. Muhuri. "A convergence theorem and an experimental study of intuitionistic fuzzy c-mean algorithm over machine learning dataset," *Applied Soft Computing*, vol. 71, pp. 1176-1188, 2018.

[23] R. Solanki, Q. M. D. Lohani, and P. K. Muhuri. "A novel clustering algorithm based on a new similarity measure over Intuitionistic fuzzy sets," In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-8. IEEE, 2015.

[24] S. Kumar, A. K. Shukla, P. K. Muhuri, and Q. M. D. Lohani. "Atanassov Intuitionistic Fuzzy Domain Adaptation to contain negative transfer learning," In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2295-2301. IEEE, 2016.

[25] R. R. Yager. "Some aspects of intuitionistic fuzzy sets," *Fuzzy Optimization and Decision Making*, vol. 8, no. 1, pp. 67-90, 2009.

[26] M. Kaushal, R. Solanki, Q. M. D. Lohani, and P. K. Muhuri. "A Novel intuitionistic fuzzy set generator with application to clustering," In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-8. IEEE, 2018.