Interval-Valued Fuzzy c-Means Algorithm and Interval-Valued Density-Based Fuzzy c-Means Algorithm

Ayush K. Varshney **Priyanka Mehra** **Pranab K. Muhuri** **Q. M. Danish Lohani**
Department of Computer Science, South Asian University, New Delhi, India
Email: varshneyayush90@gmail.com pmehra98.pm@gmail.com pranabmuhuri@cs.sau.ac.in danishlohani@cs.sau.ac.in

Abstract— **Most of the time membership value in the fuzzy set cannot be exactly defined. Interval-valued fuzzy set (IVFS) is a special type of type-2 fuzzy sets which represents the membership value of the fuzzy set as an interval. IVFS assumes that membership interval can better represent the uncertainty in the data. Accordingly, IVFS can be used to obtain good clustering results since it can represent the uncertainty more appropriately. Thus, this paper proposes the interval-valued fuzzy c-means algorithm (IVFCM) which uses IVFSs to represent the data. The concept of the proposed IVFCM is then extended to introduce the interval-valued density based fuzzy c-means (IVDFCM) algorithm based on the distance measure of IVFSs. Both IVFCM and IVDFCM are simulated over various UCI benchmark datasets to show their suitability and supremacy over their existing counterparts.**

Keywords—**Interval-valued fuzzy set, fuzzy c-means algorithm, density-based c-means algorithm, IVFS distance measure.**

I. Introduction

Clustering is an important problem in the field of Machine learning and data mining. Clustering is used to group similar objects together and dissimilar objects in a different group. Broadly, clustering can be categorized into two types: hard clustering and fuzzy clustering. Hard clustering assigns a data object into one cluster while fuzzy clustering can assign a data object into many clusters with different membership grades. Fuzzy sets (FSs) are used to represent the uncertainty that exist in the data, and a membership function (MF) of a FS is used to find the degree of belongingness of its elements to the set [1]. Most of the time, predicted membership value approximates the degree of belongingness but should not represent the exact degree. A special case of type-2 fuzzy set (T2 FS) [2], interval-valued fuzzy set (IVFS) [3] defines an interval in which membership value for an element is an interval.

Most of the fuzzy clustering approaches are based on fuzzy c-means algorithm (FCM) [4]. FCM is a fuzzy variant of the well-known k-means algorithm [5]. In the literature, various variants of the FCM algorithm were proposed such as: intuitionistic fuzzy c-means (IFCM) algorithm [6], modified fuzzy c-means algorithm [7], modified intuitionistic fuzzy c-means algorithm [8], probabilistic intuitionistic fuzzy c-means (PIFCM) algorithm [9] and many more. To represent the data, these algorithms use variants of fuzzy sets such as: ordinary FSs, intuitionistic fuzzy sets [10], interval type-2 fuzzy sets [11], type-2 fuzzy sets [2], vague sets [12], rough sets [13] and other variants of FSs [14]. These variants offer technical capabilities to represent the uncertainty in the data. Though IVFSs can represent the uncertainty of the data appropriately, very little research has been done in the field of interval-valued fuzzy clustering.

FCM is an iterative clustering algorithm. FCM initializes cluster centroids randomly and assigns data points to the clusters based on the distance measure used. It then computes the average of the data points in a cluster to find the new cluster centroids until the cluster centroids converges. FCM is highly dependent on the choice of the initialized cluster centroids. Recently proposed density based fuzzy c-means (DFCM) algorithm removes the random initialization from the FCM and finds the cluster centroid based on the density of the data points [15]. Interval-valued possibilistic fuzzy c-means (IPFCM) algorithm used the two fuzzifier constants to generate IVFSs [16]. IPFCM is an inter-valued update of the possibilistic fuzzy c-means (PFCM) algorithm [17]. There was another approach for interval-valued extension of the FCM as was proposed in [18]. Here, authors proposed interval-valued fuzzy partition with the constraint with only the lower membership values sum up to one. The interval computed in [18] for the membership, non-membership values can often be large. It also uses an adjustment parameter K which makes it computationally infeasible.

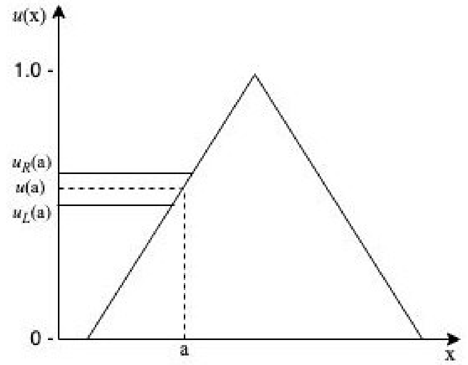

Fig. 1. Membership interval for a data value ‘a’.

This paper proposes an interval-valued fuzzy c-means (IVFCM) clustering algorithm. It uses IVFSs to represent the data and the IVFSs’ Euclidean distance measure. Instead of using two membership functions to generate the interval of membership, in this paper, we have assumed that membership function approximates the membership degree which can be used to obtain the interval for the membership value. We have taken ceiling and floor of the membership function values as the interval membership value, as shown in Fig. 1. In Fig. 1, for a data point ‘a’, u(a) is the membership value as found by using a suitable membership function for a . Here, $[u_L(\text{a}), u_R(\text{a})]$ is the interval obtained for the membership value of a, where u_L(a) is the floor of u(a) and u_R(a) is the ceiling of u(a) . Proposed IVFCM doesn't require any adjustment parameter as in [18]. It assumes that the used membership function can approximate the data points. The interval chosen has fixed length (0.1) as described above. Proposed IVFCM is a cost-effective technique for inter-valued FCM. We use the IVFSs' Euclidean distance measure to propose our algorithm. An extension of the IVFCM with recently proposed DFCM called, 'interval-valued density based fuzzy c-means (IVDFCM) Algorithm' is also proposed in this paper. IVDFCM initializes the cluster centroids based on their density. Proposed algorithms have been compared with well-known FCM and DFCM. Results have shown that proposed algorithms outperform their existing counterparts.

The major contribution of this paper is as follows:

1) A new clustering algorithm called 'interval-valued fuzzy c-means (IVFCM) algorithm' is proposed using the interval value fuzzy sets.
2) An extension of IVFCM called 'Interval-valued Density based fuzzy c-means (IVDFCM) algorithm' has also been proposed.

This paper is organized as follows: Section-II provides the required basic concepts, Section-III provides the proposed algorithms and the related explanations, Section-IV contains the experimental results, and Section-V gives a brief conclusion of the work and highlights some future directions.

II. Basic Concepts

A. Fuzzy Sets (FSs)

Let X be the universe of discourse. A Fuzzy set A for $x \in X$ can be represented as [1]:

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

where $\mu_A(x) \in [0,1]$ is the membership value for element x. It's non-membership value, $v_A(x)$ is given by:

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

B. Interval Valued Fuzzy sets

For an element x in the universe of discourse X, an interval valued fuzzy set $\tilde{A}$ for x can be defined as [3]:

$$\tilde{A} = \{\langle x, [\mu_A^L, \mu_A^R]\rangle | x \in X\} \tag{2}$$

where $[\mu_A^L, \mu_A^R]$ is an interval between $[0,1]$. For all x, $[\mu_A^L, \mu_A^R]$ are the membership interval iff:

$$0 \le \mu_A^L \le \mu_A^R \le 1. \tag{3}$$

C. Distance Measure for FSs

The distance measure for FSs $d: X \times Y \to [0,1]$ is a function which is used to compute the distance between two FSs $A\ (\in X)$ and $B\ (\in Y)$. It satisfies the following properties [19]:

$$\left.\begin{array}{ll} 1: & 0 \le d(A,B) \le 1 \\ 2: & d(A,B) = d(B,A) \\ 3: & d(A,B) = 0, \text{iff A = B} \\ 4: & \text{For, } A \subseteq B \subseteq C,\ d \text{ follows:} \\ & d(A,C) \ge d(A,B) \text{ and } d(A,C) \ge d(B,C). \end{array}\right\} \tag{4}$$

D. Normalized Euclidean Distance Measure for FSs

Szmidt and Kacprzyk [20] defined the normalized Euclidian distance Measure as follows:

$$\bar{d}(A,B) = [\tfrac{1}{2n}\textstyle\sum_{i=1}^{n}\left(\mu_A(x_i) - \mu_B(x_i)\right)^2]^{\frac{1}{2}} \tag{5}$$

E. Normalized Euclidean Distance Measure for IVFSs

Normalized Euclidean distance measure $\tilde{d}(A,B)$ for two IVFSs A and B may be defined as follows [21]:

$$\tilde{d}(A,B) = \begin{bmatrix} \left(\mu_A^L(x_i) - \mu_B^L(x_i)\right)^2 + \\ \left(\mu_A^R(x_i) - \mu_B^R(x_i)\right)^2 \end{bmatrix}^{\frac{1}{2}} \tag{6}$$

F. Fuzzy c-Means Algorithm (FCM)

Normalized Euclidian distance measure defined in Eq. (5) acts as the proximity function in FCM [4]. In FCM, FSs are used to represent the real-valued data points. FCM clusters p data points, each in n-th dimension, into c clusters. The objective function for FCM is given below:

$$\left.\begin{array}{l} \min J_m = \sum_{i=1}^{p}\sum_{j=1}^{c} u_{ij}^m \bar{d}_{ij}^2 \\ \text{s.t.} \quad \sum_{j=1}^{c} u_{ij} = 1, 1 \le j \le c \\ u_{ij} \ge 0,\ \ 1 \le i \le p, 1 \le j \le c \end{array}\right\} \tag{7}$$

$$\sum_{i=1}^{p} u_{ij} > 0,\ \ 1 \le j \le c$$

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

G. Density based Fuzzy c-Means Algorithm (DFCM) [15]:

DFCM yields good clustering results in comparison to FCM by initializing cluster centroids on the basis of density. FCM is highly sensitive to the randomly initialized cluster centroids, a drawback which DFCM removes. DFCM computes the density of a data point (ρ_i) and it's cut-off density (r_c) for determining the set of potential cluster centroids (C_p). DFCM also uses two adjustable parameters, the density rate (σ) to control the cut-off density, and the distance rate (λ) to control the distance between the potential cluster centroids. Mathematically, these parameters can be described as follows:

$$\rho_i = \textstyle\sum_j X(d_{ij} - d_c) \tag{8}$$

Here, $X: R \to \{0,1\}$ is a function over x such that,

$$X(x) = \begin{cases} 1, & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases}$$

Also, the d_{ij} function computes the Euclidian distance between the i-th and the j-th data points. The d_C is a random constant function which chooses values from [0,1]. In DFCM, all the data points are collected in a set X_s in non-increasing order of the densities. The cut-off density r_c can be derived by using the formula given below:

$$r_c = \left(\textstyle\sum_{i=1}^{p} \rho_i\right) \times \sigma \tag{9}$$

Then, r_c is used to find the set of potential clusters (C_p) as follows:

$$C_p = \left\{\tilde{x}_i \middle| \textstyle\sum_{i=1}^{k-1} \tilde{p}_i < r_c,\ \ \sum_{i=1}^{k} \tilde{p}_i \ge r_c,\ \ i = 1,2,\dots,k\right\} \tag{10}$$

Each of the cluster centroids should be some distance (δ) apart, which depends on the distance rate λ as follows:

$$\delta = \tfrac{1}{k}\left(\textstyle\sum_{i,j=1}^{k} \tilde{d}_{ij}\right) * \lambda \tag{11}$$

Based on δ, set of the potential clusters is used to initialize the cluster centroids.

DFCM also defines density-based membership function as follows:

$$\tilde{u}_{ij} = \begin{cases} 1, & \text{if } \tilde{x}_i \in V \text{ and } j = k \\ 0, & \text{if } \tilde{x}_i \in V \text{ and } j \neq k \\ \frac{\tilde{\rho}_i/\rho_{v1}}{k-1}, & \text{if } \tilde{x}_i \notin V,\ \rho_{v_k} \leq \tilde{\rho}_i \leq \rho_{v_{k-1}} \text{and } j \leq k-1 \\ \frac{1-\tilde{\rho}_i/\rho_{v1}}{c-(k-1)}, & \text{if } \tilde{x}_i \notin V,\ \rho_{v_k} \leq \tilde{\rho}_i \leq \rho_{v_{k-1}} \text{and } j \leq k-1 \end{cases} \quad (12)$$

DFCM differs from the FCM because it contains unique initialization process.

H. IVFSs generation techniques:

We assume that if a good membership function is chosen instead of using two membership function to find the membership interval, a good approximation of membership value is obtained. We have taken min-max normalization as the membership function. Floor of the function is taken as the lower membership limit, μ^L and ceiling of the membership value acts as the upper membership limit, μ^R. Therefore, chosen membership function [22,23] is as follows:

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

Hence, $\mu_A^L(x_i) = floor(\frac{x_i - x_{imin}}{x_{imax} - x_{imin}})$

and, $\mu_A^R(x_i) = ceil(\frac{x_i - x_{imin}}{x_{imax} - x_{imin}})$.

From Eq. (2), an IVFS $\tilde{A}$ can be represented as follows:

$$\tilde{A} = \{\langle x, [floor(\frac{x_i - x_{imin}}{x_{imax} - x_{imin}}), ceil(\frac{x_i - x_{imin}}{x_{imax} - x_{imin}})]\rangle \mid x \in X\} \quad (13)$$

III. Proposed Work

In this section, we introduce our proposed interval-valued fuzzy c-means (IVFCM) algorithm. We also introduce it's one extension which is interval-valued density based fuzzy c-means (IVDFCM) algorithm in this section. Eq. (13) is used to generate the IVFSs. The normalized Euclidean distance defined in the Eq. (6) is taken as the distance measure. First, we provide the flowchart of the proposed IVFCM, and then we discuss its different steps by giving its pseudocodes as *Algorithm 1*. The pseudocode of the proposed IVDFCM is given as *Algorithm 2* followed by its discussions.

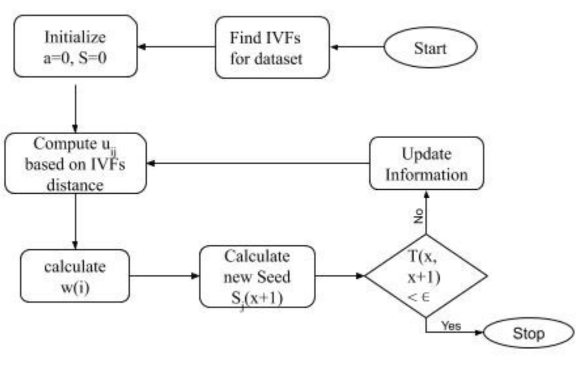

Fig. 2. Flowchart of IVFCM

As mentioned in the *Algorithm 1*, in the IVFCM algorithm, we first find the IVFSs for each of the data points using the IVFS generation technique defined in the Eq. (13). Then we use Eq. (6) as the distance measure. We then randomly choose the initial cluster centroids and find the initial partition matrix. We re-compute the cluster centroids and partition matrix at each iteration until the algorithm converges.

Algorithm 1: *Interval valued Fuzzy c-Means Algorithm*

1: Initialize c, ϵ, m, z(iteration) $= 0$ and p data points matrix A using eq. (13) and random initialized centroids S(0).
2: while (true)
3: Compute partition matrix M=$(u_{ij})_{p\times c}$ using distance measure defined in eq. (6)
4: If $\tilde{d}_2(A_i, S_z(r)) = 0$
5: $u_{ik} = 1$ and $u_{ij} = 0\ \forall\ j \neq k$.
6: Else
7: $u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\tilde{d}_2^2(A_i, S_j)}{\tilde{d}_2^2(A_i, S_k)}\right)^{\frac{2}{m-1}}}$
8: End
9: Compute new centroids $S(z+1)$ using weight defined as: $w_j = \left\{\frac{u_{ij}^m}{\sum_{i=1}^{p} u_{ij}^m},\ 1 \leq i \leq p\right\}$
10: If $\sum_{k=1}^{c} \frac{\tilde{d}_2(S_z(r), S_{z+1}(r))}{c} < \epsilon$
11: Break;
12: end while
13: END.

Algorithm 2: *Interval valued Density-based Fuzzy c-Means Algorithm*

1: Initialize λ, ϵ, m, z $= 0$ and p data points matrix A using eq. (13).
2: Initialized cluster centroids S(0) by using eq. (10), eq. (11) and section II.F.
3: while (true)
4: Compute partition matrix M=$(u_{ij})_{p\times c}$ using distance measure defined in eq. (6)
5: If $\tilde{d}_2(A_i, S_z(r)) = 0$
6: $u_{ik} = 1$ and $u_{ij} = 0\ \forall\ j \neq k$.
7: Else
8: $u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\tilde{d}_2^2(A_i, S_j)}{\tilde{d}_2^2(A_i, S_k)}\right)^{\frac{2}{m-1}}}$
9: End
10: Compute new centroids $S(z+1)$ using weight defined as: $w_j = \left\{\frac{u_{ij}^m}{\sum_{i=1}^{p} u_{ij}^m},\ 1 \leq i \leq p\right\}$
11: If $\sum_{k=1}^{c} \frac{\tilde{d}_2(S_z(r), S_{z+1}(r))}{c} < \epsilon$
12: Break;
13: end while
14: END.

We shall now discuss our proposed IVDFCM algorithm, which is shown above in its pseudocode format as *Algorithm 2*. IVDFCM algorithm also uses IVFSs to represent the data and the distance measure given in Eq. (6). It choses cluster centroids on the basis of the density of the data points. Again, we have used the min-max normalization as the membership function for the IVDFCM instead of Eq. (12), the membership function used by DFCM. The procedure for selection of the cluster centroids has been explained in the Section II.G.

Note: IVFCM and IVDFCM differs respectively from the FCM and DFCM in the representation of the data points and the used distance measure. FCM and DFCM uses the distance measures defined in Eq. (5), whereas IVFCM and IVDFCM uses the distance measure defined in the Eq. (6).

Computational Complexity of IVFCM and IVDFCM: Both of the proposed algorithms, IVFCM and IVDFCM, are independent of any adjustment parameter and uses $O(1)$ time to generate the membership interval and the non-membership interval, and hence they do not add any extra costs. The computational complexity of the IVFCM is same as that of FCM, which is $O(pnc^2i)$. The computational complexity of the IVDFCM is same as of that DFCM, which is $O(p^2 + pnc^2i)$, where p is the number of data points, d is the dimension of each data point, c is the number of clusters and i is the iteration required to converge.

IV. EXPERIMENTAL RESULTS

In this section, we will show the comparison of performance of the FCM, IVFCM, DFCM, and IVDFCM algorithms over various UCI benchmark datasets [24]. We have used three performance indices, viz., clustering accuracy, partition coefficient and cluster entropy to compare our proposed algorithms with their counterparts. Table I gives the details of the used datasets.

TABLE I. DETAILS OF BENCHMARK 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
IMAGE SEGMENTATION	2310	19	7
IRIS	150	4	3
WINE	178	13	3
ZOO	101	17	7

Clustering accuracy: Clustering accuracy is one of the highly used performance metric to compare the results of clustering algorithms. Accordingly, we have used clustering accuracy to compare the performances of the FCM, DFCM, and the proposed IVFCM and IVDFCM algorithms. Mathematically, clustering accuracy is defined as follows:

$$Clustering\ Accuracy = \frac{Nunber\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples}$$

Validation Index: Two validation indices, viz., partition coefficient (V_{PC}) and cluster entropy (V_{CE}) are used to show the superiority of our proposed algorithms over their counterparts. We may compute V_{PC} and V_{CE} as follows [25]:

$$V_{PC} = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{c} u_{ij}^2,$$

$$V_{CE} = -\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{c} u_{ij}\log u_{ij}$$

For good clustering results, we look for high clustering accuracy, high partition coefficient (V_{PC}) and low cluster entropy (V_{CE}).

(A) Comparison based on clustering accuracy

Comparison of clustering Accuracy of IVFCM, IVDFCM over various UCI benchmark datasets: Table II shows the clustering accuracies of the FCM and the DFCM algorithms over 9 UCI datasets, whereas Table III provides the clustering accuracies of the IVFCM and the IVDFCM algorithms.

TABLE II. CLUSTERING ACCURACIES OF THE FCM AND DFCM ALGORITHMS OVER VARIOUS UCI DATASETS.

Datasets (size)	FCM (%)	IVFCM (%)
IRIS	**90.67**	**90.67**
	m = 2.1	m = 1.3
ZOO	83.17	**92.08**
	m = 1.8	m = 1.6
WINE	93.25	**94.94**
	m = 3.9	m = 1.7
BREAST CANCER	92.09	**92.97**
	m = 1.2	m = 2.4
BALANCE SCALE	71.2	**72.48**
	m = 1.9	m = 1.4
IMAGE SEGMENTATION	64.81	**65.48**
	m = 2.8	m = 2.0
CAR EVALUATION	76.04	**77.78**
	m = 1.2	m = 1.1
DERMATOLOGY	**89.34**	**89.34**
	m = 2.8	m = 1.4
ECOLI	79.46	**81.85**
	m = 3.6	m = 1.1

TABLE III. CLUSTERING ACCURACIES OF THE DFCM AND IVDFCM ALGORITHMS OVER VARIOUS UCI DATASETS.

Datasets (size)	DFCM (%)	IVDFCM
IRIS	**92**	91.33
	m=1.1, dc=0.2,	m=2.4, dc=0.2,
	λ=0.25, σ=0.2	λ=0.25, σ=0.1
ZOO	93.06	**95.05**
	m=1.1, dc=0.1,	m=3.6, dc=0.4,
	λ=1.0, σ=0.5	λ=0.7, σ=0.3
WINE	96.07	**96.63**
	m=2.8, dc=0.1,	m=1.5, dc=0.2,
	λ=1.0, σ=0.2	λ=0.25, σ=0.2
BREAST CANCER	**93.67**	**93.67**
	m=3.8, dc=0.5,	m=3.4, dc=0.2,
	λ=1 .0, σ=0.4	λ=0.25, σ=1.0
BALANCE SCALE	76.64	**76.96**
	m=3.0, dc=0.1,	m=1.9, dc=0.2,
	λ=1.0, σ=0.6	λ=0.4, σ=0.1
IMAGE SEGMENTATION	77.67	**80.81**
	m=2.0, dc=0.1,	m=1.2, dc=0.2,
	λ=1.0, σ=0.2	λ=1.0, σ=0.2
CAR EVALUATION	79.46	**79.52**
	m=1.2, dc=0.2,	m=1.1, dc=0.2,
	λ=1.0, σ=0.2	λ=0.15, σ=0.2
DERMATOLOGY	**96.99**	96.18
	m=1.2, dc=0.1,	m=1.1, dc=0.4,
	λ=1.0, σ=1.0	λ=0.15, σ=0.60
ECOLI	86.41	**86.91**
	m=2.8, dc=0.1,	m=2.5, dc=0.2,
	λ=0.4, σ=0.6	λ=0.7, σ=0.6

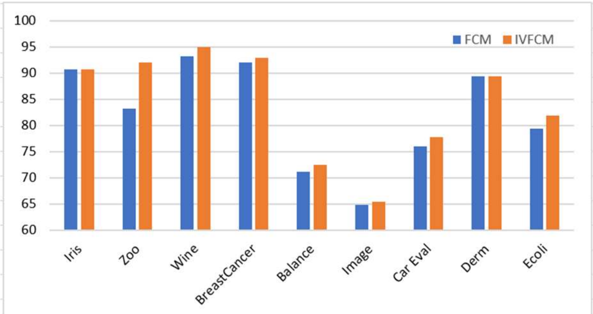

Fig.3. Comparison of FCM, IVFCM over various benchmark datasets.

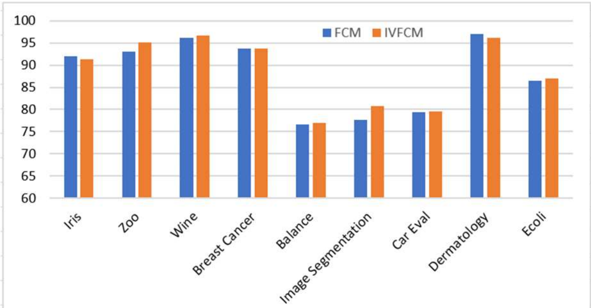

Fig.4. Comparison of DFCM, IVDFCM over various benchmark datasets.

As may be seen from the Table II, clustering accuracies of the proposed IVFCM algorithm is better than the FCM algorithm for all the considered datasets except two datasets (IRIS and DERMATALOGY). For these two datasets, both performance of both FCM and IVFCM are at par with each other, i.e., both gave equal clustering accuracies.

From the Table III, we can see that the proposed IVDFCM algorithm gives better clustering accuracies than the DFCM algorithm for most of the datasets except three data sets (IRIS and BREAST CANCER). For the IRIS and the DERMATOLOGY datasets, DFCM provided better clustering accuracies, whereas for the BREAST CANCER dataset both DFCM and IVDFCM provided equal clustering accuracies.

Fig. 3 shows the comparative pictorial depiction of the clustering accuracies of FCM against IVFCM over various benchmark datasets. Similarly, Fig. 4 visualizes the same for DFCM against the proposed IVDFCM. From both these figures, we can say that our proposed algorithms outperform their existing counterparts.

(B) Comparison based on validation indices

FCM and IVFCM: From the Table IV, we can see that the proposed IVFCM has performed better in terms of validation index over iris and dermatology datasets, even when they both resulted same clustering accuracies (*see* Table II). DFCM and IVDFCM: From Table IV, it can be seen that our proposed IVDFCM has performed better in terms of validation index over the IRIS and the DERMATOLOGY datasets, even when DFCM gave better clustering accuracies for these two datasets (*see* Table III).

TABLE IV. COMPARISON OF FCM AND IVFCM ALGORITHMS BASED ON VALIDATION INDEXES

Dataset	*Algorithm*	V_{PC}	V_{CE}
IRIS	FCM	0.880	0.213
	IVFCM	**0.986**	**0.024**
DERMATOLOGY	FCM	0.306	1.446
	IVFCM	**0.521**	**0.924**

TABLE V. COMPARISON OF DFCM AND IVDFCM ALGORITHMS BASED ON VALIDATION INDEXES

Dataset	Algorithm	V_{PC}	V_{CE}
IRIS	DFCM	0.614	0.683
	IVDFCM	**0.835**	**0.304**
DERMATOLOGY	DFCM	0.960	0.072
	IVDFCM	**0.984**	**0.030**

V. CONCLUSIONS AND FUTURE WORK

Until now, little research has been done for interval-valued fuzzy sets in clustering domain. This paper has proposed two novel IFVS based algorithms, viz., the interval-valued fuzzy c-means (IVFCM) algorithm, and the interval-valued density based fuzzy c-means (IVDFCM) algorithm. Instead of using two membership functions, we have used ceiling and floor of the membership value to generate the interval membership value. Our proposed algorithms uses Euclidean distance measure defined for IVFSs. Experimental results considering over various benchmark datasets have shown that our proposed algorithms are superior to their existing counterparts in terms of clustering accuracy and the validation indices. This paper has also highlighted the drawbacks of the proposed algorithm. It has assumed that the membership function is good at approximating the membership value for the set. It also assumed that interval required for the generation of IVFS can be taken as the ceiling and the floor of the membership value. Assumption of interval range of 0.1 may or may not work. A flexible interval range may also predict better clustering results.

A different technique for the generation of the IVFSs which can generate tighter or looser interval value depending on the dataset can be a good direction for future work. Further, recently it has been reported that IVFSs may be generalized as the interval type-2 fuzzy sets (IT2FSs) [26], [27]. The concept of IT2 FSs was floated by Mendel his co-authors in [28] based on the idea of type-2 fuzzy sets (T2 FSs) proposed by Zadeh [29]. During the last two decades many researchers have used T2 FSs and IT2 FSs for uncertainty handling in the data in many domains such as energy efficient scheduling [30], linguistic group decision making [31], clustering [32], [33], big data [34] and service quality evaluation [35]. Moreover, it was also reported by Shukla et al. [36] that research on T2 FSs, which has attracted profound attention of the scientific community, are now enough matured. Therefore, one fruitful future research work may be to extend the proposed IVDFCM for clustering the data wherein uncertainties are handled with T2 FSs or IT2 FSs, to investigate their comparative performance potentials. Similarly, intuitionistic fuzzy sets (IFSs), proposed by Atanassov [10] has also been showing promising results in many problems including clustering [37]-[41]. So, extending the algorithms proposed in this work for data clustering with IFS based uncertainty handling for a comparative study may be another prospective future research direction.

REFERENCES

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

[2] N. N. Karnik, and J. M. Mendel. "Introduction to type-2 fuzzy logic systems," In 1998 IEEE international conference on fuzzy systems proceedings. IEEE world congress on computational intelligence (Cat. No. 98CH36228), vol. 2, pp. 915-920. IEEE, 1998.

[3] W. Guijun, and L. Xiaoping. "The applications of interval-valued fuzzy numbers and interval-distribution numbers," *Fuzzy Sets and Systems*, vol. 98, no. 3, pp. 331-335, 1998.

[4] 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.

[5] M. A. Wong, and J. A. Hartigan. "A k-means clustering algorithm." *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100-108, 1979.

[6] 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.

[7] J. Kang, L. Min, Q. Luan, X Li, and J. Liu. "Novel modified fuzzy c-means algorithm with applications, " *Digital signal processing*, vol. 19, no. 2, pp. 309-319, 2009.

[8] 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.

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

[11] Q. Liang, and J. M. Mendel. "Interval type-2 fuzzy logic systems: theory and design," *IEEE Transactions on Fuzzy systems*, vol. 8, no. 5, pp. 535-550, 2000.

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

[13] D. Dubois, and H. Prade. "Rough fuzzy sets and fuzzy rough sets," *International Journal of General System*, vol. 17, no. 2-3, pp. 191-209, 1990.

[14] 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.

[15] H.-X. Pei, , Z. R. Zheng, C. Wang, C.-N. Li, and Y.-H. Shao. "D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation," *Procedia Computer Science*, vol. 122, pp. 407-414, 2017.

[16] Z. Ji, Y. Xia, Q. Sun, and G. Cao. "Interval-valued possibilistic fuzzy C-means clustering algorithm," *Fuzzy Sets and Systems*, vol. 253, pp. 138-156, 2014.

[17] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek. "A possibilistic fuzzy c-means clustering algorithm," IEEE transactions on fuzzy systems, vol. 13, no. 4, pp. 517-530, 2005.

[18] M. Pagola, A. Jurio, E. Barrenechea, J. Fernández, and H. Bustince. "Interval-valued fuzzy clustering." In *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*. Atlantis Press, 2015.

[19] L. Xuecheng. "Entropy, distance measure and similarity measure of fuzzy sets and their relations," *Fuzzy sets and systems*, vol. 52, no. 3, pp. 305-318, 1992.

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

[21] P. Grzegorzewski. "Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric," *Fuzzy sets and systems*, vol. 148, no. 2, pp. 319-328, 2004.

[22] J. Valderrama, J. F. Botía, and D. J. L. B. Valderrama. "Two cluster validity indices for the LAMDA clustering method," *Applied Soft Computing*, vol. 89, 106102, 2020.

[23] F. A. Ruiz, C. V. Isaza, A. F. Agudelo, and J. R. Agudelo. "A new criterion to validate and improve the classification process of LAMDA algorithm applied to diesel engines," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 117-127, 2017.

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

[25] N. R. Pal, and J. C. Bezdek. "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy systems*, vol. 3, no. 3, pp. 370-379, 1995.

[26] H. B. Sola, J. Fernandez, H. Hagras, F. Herrera, M. Pagola, and E. Barrenechea. "Interval type-2 fuzzy sets are generalization of interval-valued fuzzy sets: Toward a wider view on their relationship." *IEEE Transactions on Fuzzy Systems* 23, no. 5 (2014): 1876-1882.

[27] J. M. Mendel, H. Hagras, H. Bustince, and F. Herrera. "Comments on "Interval type-2 fuzzy sets are generalization of interval-valued fuzzy sets: Towards a wide view on their relationship"." *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 1, pp. 249-250, 2015.

[28] N. N. Karnik, J. M. Mendel, and Q. Liang. "Type-2 fuzzy logic systems," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 6, pp. 643–658, 1999.

[29] 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.

[30] A. K. Shukla, R. Nath, P. K. Muhuri, and Q. M. Danish Lohani. "Energy efficient multi-objective scheduling of tasks with interval type-2 fuzzy timing constraints in an Industry 4.0 ecosystem." *Engineering Applications of Artificial Intelligence* 87 (2020): 103257.

[31] R. Ureña, G. Kou, J. Wu, F. Chiclana, and E. Herrera-Viedma. "Dealing with incomplete information in linguistic group decision making by means of Interval Type-2 Fuzzy Sets." *International Journal of Intelligent Systems* 34, no. 6 (2019): 1261-1280.

[32] T. Zhang, F. Ma, D. Yue, C. Peng, and G. M. P Ohare. "Interval Type-2 fuzzy local enhancement based rough k-means clustering considering imbalanced clusters." *IEEE Transactions on Fuzzy Systems* (2019).

[33] A. K. Shukla, and Pranab K. Muhuri. "Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets." *Engineering Applications of Artificial Intelligence* 77 (2019): 268-282.

[34] A. K. Shukla, M. Yadav, S. Kumar, and P. K. Muhuri. "Veracity handling and instance reduction in big data using interval type-2 fuzzy sets." *Engineering Applications of Artificial Intelligence* 88 (2020): 103315.

[35] G. Tang, F. Chiclana, X. C. Lin, and P. Liu. "Interval type-2 fuzzy multi-attribute decision-making approaches for evaluating the service quality of Chinese commercial banks." *Knowledge-Based Systems* (2020): 105438.

[36] A. K. Shukla, S. K. Banshal, T. Seth, A. Basu, R. John, and P. K. Muhuri. "A Bibliometric Overview of the Field of Type-2 Fuzzy Sets and Systems [Discussion Forum]." *IEEE Computational Intelligence Magazine* 15, no. 1 (2020): 89-98.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.