ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2017.0213

AN APPLICATION OF PSO-BASED INTUITIONISTIC FUZZY CLUSTERING TO MEDICAL DATASETS

S. Parvathavarthini[1], N. Karthikeyani Visalakshi[2], S. Shanthi[3] and K. Lakshmi[4]
[1,3,4]School of Computer Technology and Applications, Kongu Engineering College, India
[2]Department of Computer Science, NKR Government Arts College for Women, India

Abstract

Clustering is the process of splitting data into several groups based on the characteristics of data. Fuzzy clustering assigns a data object to various clusters based on different membership values. In medical field, the diagnosis of the disease has to be done without faults and in an earlier time without any delay. So, there is a need to represent imprecise nature of the data. To represent vague data in a clear manner, Intuitionistic fuzzy set introduces a parameter called hesitancy degree. In case of Intuitionistic fuzzy clustering, this indicates that the user is not aware whether the object belongs to or not belongs to a cluster. In such a case, hesitancy can very well represent the inherent noise in the data or the ignorance of the user that is given by the state 'may be'. All clustering algorithms choose the initial seed in a random fashion. But, this creates a serious impact on the convergence of the algorithm and the clustering algorithms tend to fall into local minima. This work utilizes Intuitionistic fuzzy Particle Swarm Optimization to initialize the centroids for the Intuitionistic fuzzy clustering algorithm. The algorithm is executed over medical datasets from UCI repository and the results indicate that optimal clusters are achieved. The proposed method performs well when compared with IFCM and FCM-PSO.

Keywords

Clustering, Intuitionistic Fuzzy Set, Particle Swarm Optimization, Inertia weight, Lambda value

1. INTRODUCTION

Data Mining pertains to the task of discovering hidden knowledge from a huge volume of data. Data Mining recognizes the patterns that are available in data with the help of several techniques like Classification, Clustering, Association rule mining, Prediction, etc. Classification is a supervised technique that categorizes data as belonging to which class. Prediction tries to guess the relationship between the variables in data objects and Association rule mining correlates the behavior of data with the outcome of events. Data Mining finds its applications in various fields like Biomedical research, Behavioral and social sciences, Earth sciences, Market Analysis, web search, Decision Support Systems, Buying pattern prediction, etc.

Nowadays, voluminous data is available in all fields. It is very difficult to handle and analyze all these data manually. Clustering helps in effective decision making that can be applied to various fields like business intelligence, social media analysis, medical diagnosis, opinion analysis, satellite image segmentation, etc.

Clustering segregates data into several grou ps based on their traits. Clustering algorithms can be classified as hard or soft. Hard clustering algorithms allocate an object to exactly one cluster. Soft clustering allows an object to be a part of different clusters with different membership values.

Fuzzy clustering indicates a 'yes' or 'no' state only. But Intuitionistic fuzzy clustering allows another intermediate state 'may be'. Many real world clustering problems posess uncertainty as a key challenge. Thus the indeterminancy present in the data can be well depicted by Intutionistic fuzzy sets.

The problem with Fuzzy C-Means (FCM) [1] and Intuitionistic Fuzzy C-Means (IFCM) algorithms is that they tend to fall into local minima. As a result, the execution time increases and the quality of cluster structure is affected. So, an optimization algorithm can be used to select the initial seed and to reach the global optimal solution. An optimization algorithm aims at minimizing (in case of clustering) or maximizing an objective function subject to certain constraints. The role of objective function is to validate the clustering output and direct it through the optimal cluster centroids.

Particle Swarm Optimization (PSO) [2] is a renowned conventional technique that imitates the bird flocking behavior and uses two parameters called velocity and position which represent the speed with which the particle travels and the resulting change in the particle's position respectively. PSO needs some parameters to be tuned such as inertia weight, social and cognitive components dimension and rane of particles, etc.

While a particle is developing a new situation, both the cognitive component of the relative particle and the social component generated by the swarm are used. This situation facilitates the PSO algorithm to effectively make a drive away from the local solutions and move towards global optimum solutions. The ability to effectively solve highly nonlinear problems makes PSO suitable for solving high-dimensional clustering problems. This work combines PSO with IFCM to achieve rapid convergence and maximize the cluster quality.

In order to efficiently cluster real-time datasets, our contributions include

- Developing a novel, highly scalable hybrid algorithm by combining Particle Swarm Optimization with Intuitionistic Fuzzy C-Means clustering
- Combining the best features of IFCM algorithms proposed by Tchaira [16] and Xu [19]
- Proposed method thrives for Intuitiveness and ease of implementation
- Due to stochastic behavior of PSO, the global optimum results are achieved.
- Ability to deal with noisy data and produce good results in terms of objective function and cluster indices.

This paper is organized as follows: Section 2 reviews the similar works existing in the literature, section 3 gives an overview of fuzzy set and Intuitionistic fuzzy sets, section 4 focuses on IFCM clustering, section 5 throws light on PSO, section 6 explains the proposed IFPSO_IFCM algorithm and section 7 provides the experimental results and discussion.

2. LITERATURE SURVEY

PSO enables rapid searching and leads to fast convergence of the clustering algorithm. There are only a few numbers of works that have combined PSO with Intuitionistic Fuzzy (IF) clustering. Most of the researchers have utilized PSO for initializing the FCM algorithm and for segmentation of images.

Kumutha et al. [3] used Intuitionistic Fuzzy (IF) PSO to cluster gene expression datasets to yield faster convergence and reduce the complexity of IFCM. Nanda et al. [4] automatically identified the number of clusters in the dataset by combining cloning technique with PSO. Binu [5] compares PSO, Genetic Algorithm and Cuckoo search over seven newly defined objective functions and found that PSO works well for large scale data.

Izakian et al. [6] combined fuzzy PSO with FCM to minimize the objective function leading to a global solution. Benaichouche et al. [7] segmented images by considering the geometrical shape of clusters found by incorporating spatial information and Mahalanobis distance. The resulting image is reclustered using a local criterion optimization using greedy algorithm to detect the misclassified pixels.

Izakian et al. [8] utilized Fuzzy C Means and Particle Swarm Optimization to cluster moving objects or trajectories. Data is represented using discrete cosine transform.

Salmeron et al. [9] created a decision support tool for diagnosing and treating arthritis using the concept of fuzzy cognitive maps. Hebbian-based FCM learning is adapted and hybridized with PSO to calculate the severity of the disease.

Saxena et al. [10] reviewed the different methods for clustering, along with the measures for finding similarity the evaluation criteria and discussed the applications of clustering in various domains. Hein et al. [11] developed a fuzzy particle swarm reinforcement learning to create self-organized fuzzy controllers and applied it to benchmark datasets.

Nobile et al. [12] proposed Fuzzy self tuning PSO to control the parameters of PSO using fuzzy logic. Since the performance of PSO is highly dependent on these parameters like cognitive and social factors, inertia weight, upper and lower bounds of velocity, this work automatically tunes the parameters based on fuzzy rules. The efficiency of the algorithm is proved by testing it against twelve benchmark functions.

Oliveira et al. [13] presented a homogeneous cluster ensemble based on particle swarm clustering algorithm. Initially, many base partitions are taken from the data and they are given as input to the consensus function and genetic selection operators are used to decide the final partition. The algorithm is experimented over real and synthetic datasets. It eliminates the process of cluster alignment and allows for combination of partitions with different clusters.

Silva et al. [14] dynamically varied the parameters of PSO like c_1, c_2 and inertia weight during execution and proposed improved self-adaptive PSO for clustering data by reducing the number of parameters to be tuned. Mekhmoukh et al [15] used PSO to reduce the sensitivity to noise by incorporating spatial infor mation into Kernel Possibilistic C Means algorithm.

Chaira [16] developed a multi-objective criterion function for segmenting brain CT images by including hesitancy factor in the updation of cluster centers. Shanthi et al [17] utilized this clustering to classify mammogram images and built decision tree for effective diagnosis. Chaira [18] also utilized IF divergence for edge detection of Tumor/ hemorrhage regions. Xu et al. [19] applied a new method for clustering numerical data like car market data, supplier data and building materials data using Lagrange multiplier method and introduced a weighted average operator to assign weights for each IFS.

Prabhjot kaur et al. [20] presented a robust IFCM and kernel version of IFCM with a new distance metric incorporating the distance variation of data-points within each cluster. Rohan Bhargava et al [21] hybridized rough set with IFS in order to describe a cluster by its centroid and its lower and upper approximations.

Balasubramaniam [22] segmented nutrition deficiency in incomplete crop images using IFCM. The missing pixels in the incomplete images were imputed using IFCM algorithm. V.P. Ananthi et al. [23] segmented gray scale images using IFS. The entropy is calculated to find the threshold. The value that minimizes the entropy is taken as the threshold for segmenting the image.

Parvathavarthini et al. [24] combined IFCM with cuckoo search and applied it to several realtime datasets for validating the cluster structure using varius cluster indices. Parvathavarthini et al. [25] hybridized IFCM with crow search opptimization producing very low error rates for realtime datasets.

Many researchers [3][4][5][6][14][15] have proved that PSO suits well for obtaining global optimal solutions because of its intuitiveness, ease of implementation, and the ability to effectively solve highly nonlinear problems. There are very few studies regarding the hybridization of IFCM with IFPSO. This work highlights the efficiency of IFPSO_IFCM on medical datasets in terms of objective function and validity indices.

3. BACKGROUND

3.1 FUZZY SET AND INTUITIONISTIC FUZZY SET

Fuzzy sets are designed to manipulate data and information possessing non-statistical uncertainties [24]. A fuzzy set is represented by Zadeh [19] as follows,

$$FS = \left\{ \left\langle x, \mu_{FS}\left(x\right) \right\rangle \middle| x \in X \right\}$$

where, μ_{FS}: $X \rightarrow [0, 1]$ and ν_{FS}: $X \rightarrow [0, 1]$ and $\nu_{FS}(x)=1 - \mu_{FS}(x)$. Here μ_{FS} is the membership value and ν_{FS} is the non-membership value.

An Intuitionistic Fuzzy Set proposed by Atanassov [26] can be symbolized as below

$$IFS = \left\{ \left\langle x, \mu_{IF}\left(x\right), \nu_{IF}\left(x\right) \right\rangle \middle| x \in X \right\}$$

where, μ_{IF}: $X \rightarrow [0, 1]$ and ν_{IF}: $X \rightarrow [0, 1]$ define the degree of membership and non-membership, respectively and

$$\pi_{IF}(x) = 1 - \mu_{IF}(x) - \nu_{IF}(x) \text{ such that } 0 < \mu_{IF}(x) + \nu_{IF}(x) < 1$$

where, π_{IF} is the hesitancy value used to represent the uncertainty.

3.2 INTUITIONISTIC FUZZY C-MEANS CLUSTERING

The first task for IFCM algorithm [16] is to convert crisp data into fuzzy data which in turn would be converted to Intuitionistic fuzzy data. This process involves the task of fixing the lambda value which is a value that varies for each dataset. The value of lambda is chosen as the one which maximizes the entropy value. Entropy [27] is the amount of fuzziness present in any given dataset and it is calculated as,

$$IFE = \frac{1}{N \times M} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \frac{2\mu_i\left(d_j\right)\nu_i\left(d_j\right) + \pi_i^2\left(d_j\right)}{\pi_i^2\left(d_j\right) + \mu_i^2\left(d_j\right) + \nu_i^2\left(d_j\right)} \tag{1}$$

where, N and M are the rows and columns of the dataset.

The crisp data is converted into fuzzy data using the following Eq.(2)

$$\mu_i\left(d_j\right) = \frac{d_{ij} - \min\left(d_j\right)}{\max\left(d_j\right) - \min\left(d_j\right)} \tag{2}$$

where, d_{ij} is the current cell of the matrix under consideration and $\min(d_{ij})$ indicates the minimum value in the dataset matrix and $\max(d_{ij})$ indicates the maximum value in the dataset matrix.

Then the fuzzy data is converted to Intuitionistic fuzzy data as follows:

$$\mu_i\left(d_j;\lambda\right) = 1 - \left(1 - \mu_i\left(d_j\right)\right)^{\lambda} \tag{3}$$

$$\nu_i\left(d_j;\lambda\right) = 1 - \left(1 - \mu_i\left(d_j\right)\right)^{\lambda(\lambda+1)} \tag{4}$$

where, $\lambda \in [0,1]$

The intuitionistic fuzzification converts the intermediate fuzzy dataset to intuitionistic fuzzy dataset. The hesitancy factor is calculated by summing up the membership and non-membership degrees and subtracting the sum from one.

The clustering procedure given by [19] is followed. The distance matrix is calculated based on the Intuitionistic fuzzy Euclidean distance. Then, the membership matrix is calculated as follows

$$U_{ij} = \frac{1}{\sum_{r=1}^{C} \left(\frac{dis\left(d_j^{'}, v_i\right)}{dis\left(d_j^{'}, v_r\right)} \right)^{\frac{2}{m-1}}}, 1 \le i \le C,\ 1 \le j \le n,\ m = 2 \tag{5}$$

where, C is the number of clusters, n is the number of instances and m is the fuzziness parameter.

This membership value is used to calculate non-membership and hesitancy values. Using these values, the mass (weight) factor given to each attribute t is calculated as follows,

$$ma_i\left(k+1\right) = \left\{ \frac{u_{i1}\left(k\right)}{\sum_{j=1}^{n} u_{ij}\left(k\right)}, \frac{u_{i2}\left(k\right)}{\sum_{j=1}^{n} u_{ij}\left(k\right)}, \ldots, \frac{u_{in}\left(k\right)}{\sum_{j=1}^{n} u_{ij}\left(k\right)} \right\}, 1 \le i \le C \tag{6}$$

where, k indicates the previous iteration.

Using these mass values, the new centroids are calculated as

$$V_i = \left\{ \left[d_{s,} \sum_{j=1}^{n} ma_j \mu_{Aj}\left(d_s\right), \sum_{j=1}^{n} ma_j \nu_{Aj}\left(d_s\right) \right], 1 \le s \le n \right\}, 1 \le i \le C \tag{7}$$

where, d_s is the attribute value in the original dataset.

The objective function of IFCM can be given as

$$J_m(x,y) = \sum_{i=1}^{c} \sum_{j=1}^{n} U_{ij}^{m} X_j^{'} - C_i,\ 1 \le m \le \infty \tag{8}$$

where, U_{ij} indicates the membership matrix and the term ||..|| denotes the distance matrix.

This objective function should be minimized so that the bondage between the objects of same cluster is high and the inter-cluster distance between objects of various clusters is low. The iterations are continued till two consecutive iterations produce same value for the objective function.

3.3 PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) [2][28] is a population-based stochastic optimization technique inspired by bird flocking and fish schooling which is based on iterations/generations. Each particle has an initial position and it moves towards a better position with a velocity. The positions represent the solutions for the problem. Initially, the position and velocity matrices are assigned random values.

Consider the population or swarm size as m and the particle dimension as n. Let velocity be represented as $Velo_i = \{v_1, v_2, \ldots, v_n\}$ and position be represented as $Xpos_i = \{x_1, x_2, \ldots, x_n\}$ where i = 1 to n. For every iteration, these two vectors are updated using the following Eq.(9)-Eq.(10).

$$Velo(k+1) = wt \cdot Velo(k) + (c_1 \cdot rand_1) \cdot (p_{best}(k) - Xpos(k)) + (c_2 \cdot rand_2) \cdot (g_{best}(k) - Xpos(k)) \tag{9}$$

$$Xpos(k+1) = Xpos(k) + Velo(k+1) \tag{10}$$

where, c_1 and c_2 are user-defined constants, wt denotes the inertia weight, $rand_1$ and $rand_2$ are the random values from 0 to 1. The fitness is evaluated by calculating the objective function for each particle in the swarm.

The individual best performance is termed as p_{best} and it is updated by comparing fitness values of each iteration with that of the previous iteration. The overall best position attained by any particle with the overall minimum fitness (in case of minimization problems like clustering) is chosen as the g_{best}. The inspiring feature of PSO is that it exempts the possibility of the solution getting stuck in the local optima and tries to reach the global optima by converging in less number of iterations.

4. PROPOSED METHODOLOGY: IFPSO_IFCM

All the existing approaches work well for datasets which do not possess any noise. The need for Intuitionistic fuzzy clustering to be combined with Intuitionistic Particle Swarm Optimization comes into picture when there are abnormalities in the features of a data. This abnormality or error factor can be very well represented as the hesitancy value in IFS. This results in a consistent state of the particle's position.

4.1 ALGORITHM IFPSO-IFCM

Step 1: Initialize the parameters like population size, c_1, c_2, inertia weight and the maximum number of iterations, the number of clusters *C*, the problem dimension *D* and the fuzziness parameter *m*

Step 2: Convert data into IFS representation using Eq.(1), Eq.(4) and Eq.(5)

Step 3: For IFS conversion, fix the parameter lambda using Eq.(3). The lambda value which maximizes the entropy is fixed for each dataset

Step 4: Create a swarm with *P* particles

Step 5: Initialize the position x_{pos}, velocity *velo*, p_{best} and g_{best} as $n \times c$ matrices

Step 6: For each particle, compute the distance measure and thus calculate membership values of each object to various clusters using Eq.(6).

Step 7: Evaluate the fitness of each particle using Eq.(9).

Step 8: Calculate the personal best value p_{best} for each particle and the overall best performance g_{best} for the entire swarm

Step 9: Update the particle velocity and position using Eq.(10) and Eq.(11) respectively

Step 10: Repeat Step 6 to Step 9 until IFPSO converges i.e. g_{best} attains stability

Step 11: Obtain the particle that has the global best value with minimum cost and keep it as the initial set of centroids for the execution of the IFCM algorithm

Step 12: Compute the membership values using Eq.(6)

Step 13: In order to update the centroids, a mass is to be calculated for each attribute in the dataset using Eq.(7)

Step 14: As a function of mass, the centroids are updated using Eq.(8)

Step 15: Evaluate the fitness using Eq.(9)

Step 16: Repeat steps 12 to 15 until IFCM converges i.e. until the objective function converges

Step 17: If IFPSO_IFCM has met the stopping criterion to reach maximum iterations, then stop. Otherwise, go to Step 6.

Step 18: Find the index value of the cluster for each object. The cluster center which has the maximal membership will be the corresponding index.

Table.1. Parameters for IFPSO_IFCM

Parameter	Value
Fuzziness parameter	$m = 2$
Lambda	0 to 1 (based on the value that maximizes entropy)
Mass vector	$1/n$, where n is the number of attributes in dataset
Population	10
Max Iterations	100
Algorithm specific parameters	$c_1 = c_2 = 1.4$, $wt = 0.72$

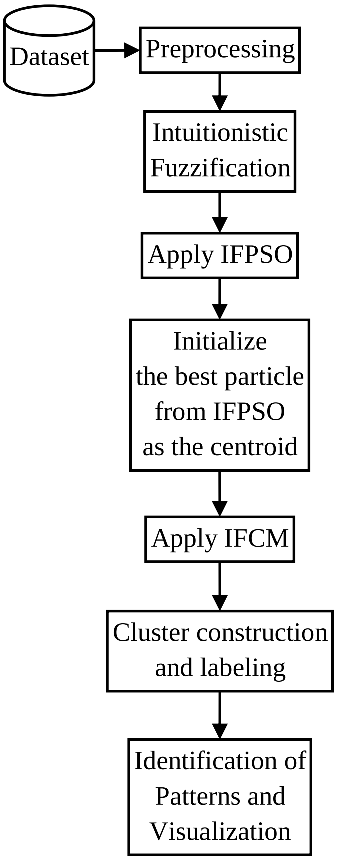
Dataset
Preprocessing
Intuitionistic Fuzzification
Apply IFPSO
Initialize the best particle from IFPSO as the centroid
Apply IFCM
Cluster construction and labeling
Identification of Patterns and Visualization

Fig.1. Workflow of the Proposed System

4.2 PSEUDOCODE FOR IFPSO-IFCM

```
Create intuitionistic fuzzy representation of data
Create intuitionistic fuzzy representation of data
Initialize the population of $N$ particles, $C$ clusters, inertia weight, constants $c_1$ and $c_2$ and maximum iterations $it_{max}$
Initialize the position and velocity of particles randomly with $N \times D$ dimension search space
While $run < max\_runs$
  While $t < it_{max}$
    Repeat
    For $A = 1:N$
      Calculate membership matrix using Eq.(5)
      Calculate fitness of each particle using Eq.(8)
      Set $p_{best}$ of each particle and $g_{best}$ of the swarm
      Update the particle velocity and position using Eq.(10) and Eq.(11) respectively
    End for
    Until $g_{best}$ converges
```

```
            Set the particle with $g_{best}$ as initial centroid
          Repeat
          Calculate membership matrix using Eq.(5)
          Calculate mass values using Eq.(6)
          Update cluster centers using Eq.(7)
          Until IFCM converges
      **End while**
  Return the cluster indices
  **End while**
```

5. EXPERIMENTAL ANALYSIS

The algorithm is implemented using MATLAB to quantitatively evaluate the performance of the proposed algorithm and the results are compared with FCM-PSO and IFCM algorithms. Experiments are conducted in two aspects: the first one with respect to the objective function value and the second one with respect to the validity indices namely the Rand Index and DB index.

Cluster validation is the predominant way of judging the performance of a clustering algorithm. Rand index and F-Measure are external validity measures and DB index is an internal measure. A greater value closer to one indicates good performance in Rand index and F-Measure. Lesser value results in good clusters in case of DB index.

Six medical datasets from UCI data repository [29] are considered for evaluating the performance. The datasets include Breast tissue, Bupa liver disorders, Contraceptive Method Choice (CMC), Dermatology, Haberman survival and Wisconsin Breast Cancer (WBC). Breast tissue dataset has 6 clusters, 9 attributes and 106 instances. It has the electrical impedance measurements of tissue samples from breast. Liver disorder has 2 clusters, 7 attributes and 345 instances. This contains blood test reports of male who had liver disorders due to excessive alcohol consumption. CMC dataset has 3 clusters, 9 attributes and 1473 instances. The samples are taken from married women to predict their contraceptive method choice based on their demographic and socio-economic characteristics.

Dermatology dataset has 6 clusters, 34 attributes and 366 instances. This dataset contains 12 clinical attributes and 22 histo-pathological attributes to identify a variety of skin diseases. Haberman dataset has 2 clusters, 3 attributes and 306 instances. This dataset contains the information on survival of patients who had undergone surgery for breast cancer. WBC dataset has 2 clusters, 32 attributes and 569 instances. This dataset is used to identify the patients with small clumps in breast as benign or malignant.

The Table.2 shows the fitness values obtained as a result of the proposed method and compares it with the IFCM and FCM-PSO algorithms. It is evident from the Table.2 that the proposed IFPSO-IFCM algorithm gives an overwhelming response in terms of the fitness values for all the six datasets. The IFCM algorithm produces a high value for all the datasets and takes more time to converge. Also, only local optimum solutions are achieved in many cases. But PSO is utilized in the other two methods for rapid searching of the optimal solution. By exploiting both the cognitive component of the relative particle and the social component generated by the swarm, PSO can reach the global optimum solutions.

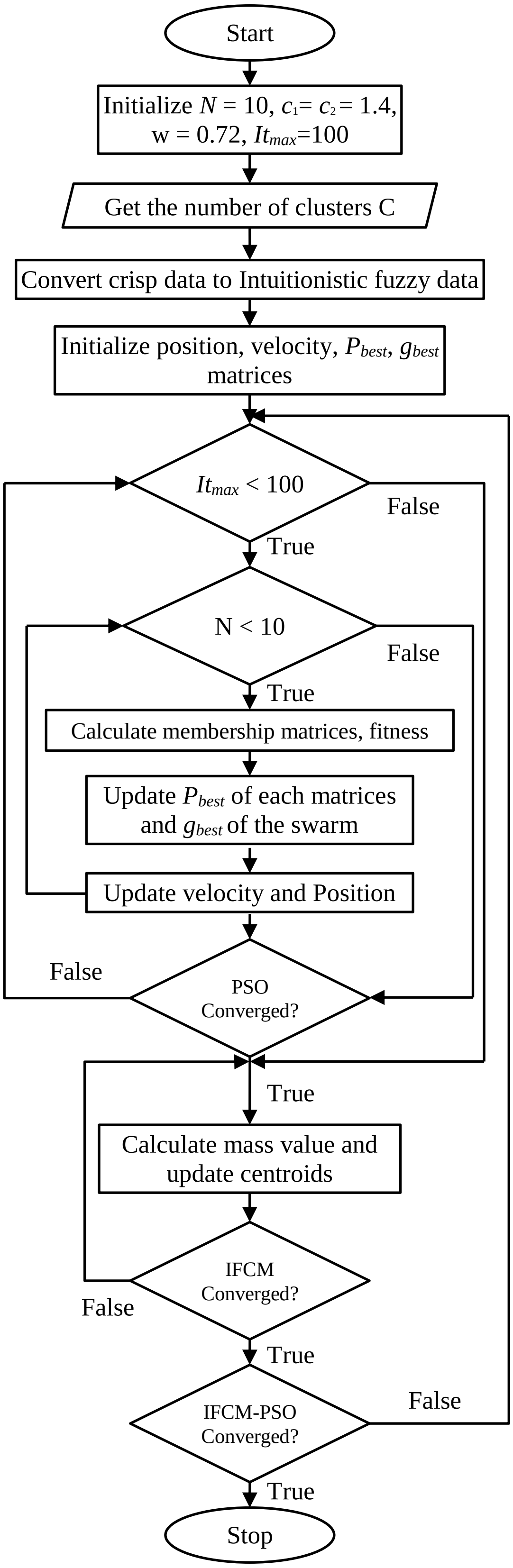
Start
Initialize N = 10, c_1= c_2= 1.4, w = 0.72, It_{max}=100
Get the number of clusters C
Convert crisp data to Intuitionistic fuzzy data
Initialize position, velocity, P_{best}, g_{best} matrices
It_{max} < 100
False
True
N < 10
False
True
Calculate membership matrices, fitness
Update P_{best} of each matrices and g_{best} of the swarm
Update velocity and Position
False
PSO Converged?
True
Calculate mass value and update centroids
IFCM Converged?
False
True
IFCM-PSO Converged?
False
True
Stop

Fig.2. Flowchart for IFPSO-IFCM

The datasets have different scales with respect to their variables. Generally, Euclidean distance is sensitive to this variation in scales and this difference can be eliminated by normalizing the variables in the range 0 to 1. Due to the fact that PSO algorithm maintains its stochastic behavior capacity, it provides high quality solutions.

Table.2. Comparison of Objective Function Values

Dataset	Values	IFCM	FCM-PSO	IFPSO-IFCM
Breast tissue	Mincost	3.15	1.87	**0.55**
	Maxcost	4.25	1.94	**0.60**
	Avgcost	3.21	1.91	**0.59**
Bupa liver disorders	Mincost	26.81	9.36	**8.61**
	Maxcost	27.06	9.83	**8.99**
	Avgcost	26.92	9.38	**8.77**
CMC	Mincost	175.18	112.5	**71.19**
	Maxcost	224.17	124.2	**86.99**
	Avgcost	183.21	113.1	**72.44**
Dermatology	Mincost	131.23	119.25	**108.12**
	Maxcost	174.54	**121.38**	128.15
	Avgcost	132.35	120.11	**113.41**
Haberman survival	Mincost	39.07	8.95	**6.88**
	Maxcost	51.01	9.45	**7.42**
	Avgcost	40.26	8.98	**6.90**
Wisconsin Breast Cancer	Mincost	58.62	21.29	**13.56**
	Maxcost	76.10	24.60	**15.19**
	Avgcost	60.76	22.58	**14.82**

The Table.2 shows that the fitness value for the datasets with more than two clusters has reduced to a great extent. The Breast tissue, CMC and Dermatology datasets have a major deviation in their objective function indicating that the proposed method works well even with more number of clusters. In case of other datasets like Bupa, Haberman and WBC, there is a significant reduction in the fitness value.

5.1 RAND INDEX

The rand index considers a set of quadruples namely true positive, true negative, false positive and false negative. True positive (TP) decision allocates two objects with similar characteristics to the same cluster whereas a true negative (TN) assigns objects with various traits to different clusters. There are two types of errors we can commit. A False positive (FP) decision assigns two dissimilar documents to the same cluster. A False negative (FN) decision assigns two similar documents to different clusters. The Rand index [30] measures the percentage of decisions that are correct. Rand Index can be calculated using the following formula

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{11}$$

The Table.3 shows the Rand Index values for the six datasets. It can be noticed that the highest Rand index value is obtained for Wisconsin Breast Cancer (WBC) dataset as 0.8123 and the least value is for liver disorder dataset. Breast tissue and dermatology, even with more number of clusters, have the next greater values. A close performance between IFCM and FCMPSO can be observed.

5.2 DAVIS-BOULDIN INDEX

The Davis-Bouldin index [31] is based on a ratio of within cluster and between cluster distances. This shows good performance when the value is less. The formula for DB Index can be given as,

$$\frac{1}{k}\sum_{i=1}^{k}\max_{j}\frac{s(C_i)+s(C_j)}{d_C(C_i,C_j)} \tag{12}$$

where, k is the number of clusters, $s(c)$ is the average distance among the instances in cluster C, $d_C(C_i,C_j)$ measures the distance between the centers of C_i and C_j.

The Table.4 shows the DB index values for the six datasets. In case of DB index, the best value is obtained again for WBC and the least value is for Haberman survival dataset. The algorithm is stable for various number of clusters and small, medium and large datasets in terms of number of instances or number of attributes. Liver disorder, WBC and dermatology datasets have values closer to zero indicating a higher level of performance. For the other datasets, a nominal value is obtained.

Table.3. Comparison of Rand Index values

Algorithm	IFPSO-IFCM	IFCM	FCMPSO
Dataset	Rand Index		
Breast tissue	0.7461	0.7269	0.7218
Liver disorders	0.6026	0.5031	0.5583
CMC	0.6457	0.5637	0.5812
Dermatology	0.7421	0.6560	0.6976
Haberman survival	0.6127	0.5128	0.6003
WBC	0.8123	0.7512	0.7994

Table.4. Comparison of DB Index values

Algorithm	IFPSO-IFCM	IFCM	FCM PSO
Dataset	**DB Index**		
Breast tissue	0.3629	0.3624	0.3724
Liver disorders	0.1315	0.2971	0.1882
CMC	0.3178	0.3216	0.4113
Dermatology	0.1207	0.3979	0.2095
Haberman survival	0.3767	0.5586	0.3942
WBC	0.0145	0.2952	0.1094

5.3 F-MEASURE

The F-Measure [32] is an external index. It is the harmonic mean of the precision and recall coefficients. If the precision is high and recall value is low, this results in a low F-measure. If both precision and recall are low, a low F-measure is obtained. On the other hand, if both are high, a high F-measure value is obtained. F-Measure can be computed using the formula given in Eq.(13),

$$F = \frac{2TP}{2TP + FP + TN} \tag{13}$$

Table.5. Comparison of F-Measure values

Algorithm	**IFPSO-IFCM**	**IFCM**	**FCMPSO**
Dataset	**F-Measure**		
Breast tissue	0.8123	0.6027	0.7589
Liver disorders	0.6915	0.6003	0.6191
CMC	0.7589	0.6121	0.5196
Dermatology	0.6394	0.5933	0.5884
Haberman survival	0.6842	0.6754	0.6775
WBC	0.8672	0.8488	0.8606

The results for F-measure values are given in Table.5. It is evident that the values for IFPSO-IFCM are significant when compared to other two algorithms. IFCM may lead to a local optima, FCMPSO utilizes optimization and almost achieves the same result for most of the datasets with additional complexity. But IFPSO-IFCM avoids local optimal solutions and also produces high quality clusters in terms of cluster validity indices. As with Rand index, WBC provides the highest F-measure value. This indicates that even with a high number of attributes (39 attributes per record), the proposed method provides efficient results after careful analysis. The CMC dataset that has greater number of instances (1473 records) achieved sufficient values for all the three indices leading to the conclusion that the algorithm IFPSO-IFCM explored and exploited the problem space in an appreciable way. Also, the breast tissue and dermatology datasets have six clusters each and even when the number of clusters increase, the algorithm exhibits a good cluster quality.

The results of the tests lead to the conclusion that IFPSO-IFCM is really better than the other two algorithms. PSO is also capable of memorizing the solutions. This helps in retaining the best individuals. The Fig.2 shows the comparative results of IFPSO-IFCM, IFCM and FCM-PSO for the Rand Index, Fig.3 compares the DB Index values obtained and Fig.4 compares the F-Measure values. The proposed methodology shows a superior performance for all the datasets.

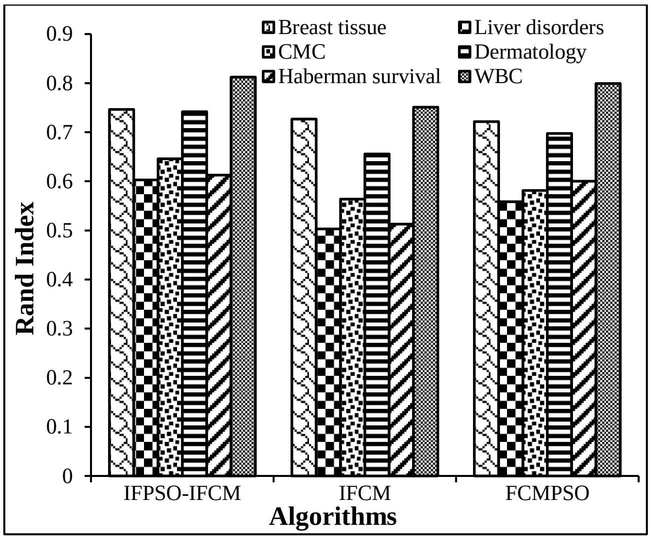

Fig.2. Rand Index Comparison

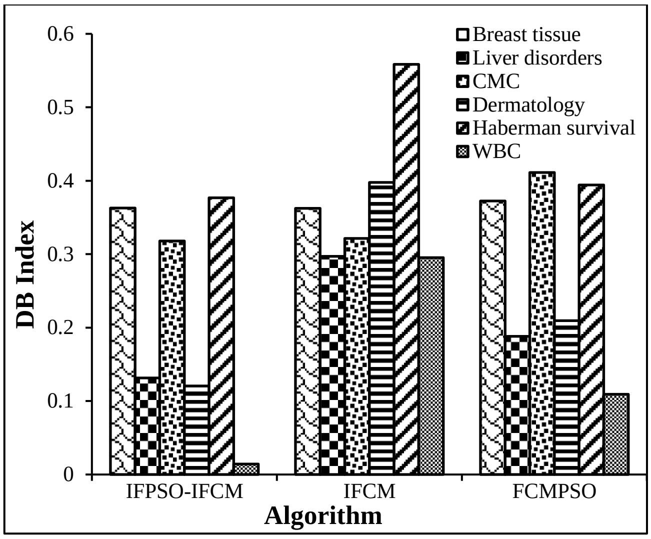

Fig.3. DB Index Comparison

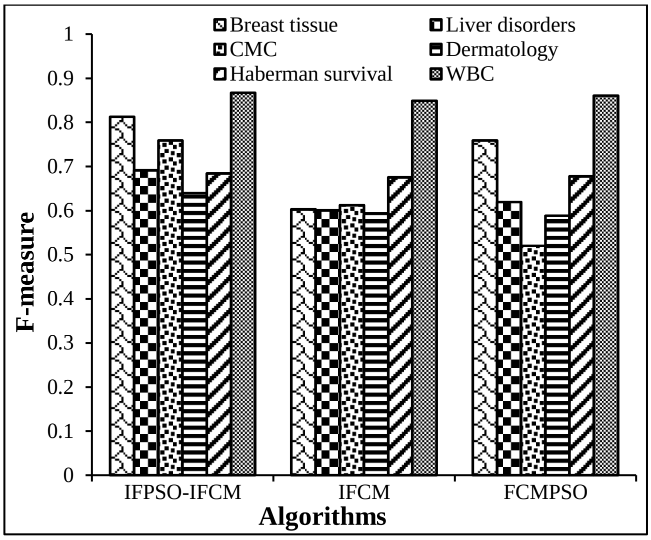

Fig.4. F-Measure Comparison

6. CONCLUSION

The FCM and IFCM algorithms tend to fall into local minima and also the convergence is delayed due to random selection of initial seeds. The IFCM algorithm is hybridized with PSO which is based on intelligence in this work. This method results in fast convergence to the sub optimal solution. Also, the performance of the algorithm is evaluated in terms of fitness function and validity indices. The results prove that the IFPSO-IFCM converges to a minimum objective function value and efficient cluster structures are obtained.

REFERENCES

[1] J.C. Bezdek, R. Ehrlich and W. Full, “FCM: The Fuzzy C-means Clustering Algorithm”, *Computers and Geosciences*, Vol. 10, No. 2-3, pp. 191-203, 1984.

[2] J. Kennedy and R. Eberhart, “Particle Swarm Optimization”, *Proceedings of IEEE International Conference Proceedings on Neural Networks*, pp. 1942-1948, 1995.

[3] V. Kumutha and S. Palaniammal, "Improved Fuzzy Clustering Method Based On Intuitionistic Fuzzy Particle Swarm Optimization", *Journal of Theoretical and Applied Information Technology*, Vol. 62, No. 1, pp. 8-15, 2014.

[4] S.J. Nanda and G. Panda, "A Survey on Nature Inspired Metaheuristic Algorithms for Partitional Clustering", *Swarm and Evolutionary Computation*, Vol. 16, pp. 1-18, 2014.

[5] D. Binu, "Cluster Analysis using Optimization Algorithms with Newly Designed Objective Functions", *Expert Systems with Applications*, Vol. 42, No. 14, pp. 5848-5859, 2015.

[6] H. Izakian and A. Abraham, "Fuzzy C-means and Fuzzy Swarm for Fuzzy Clustering Problem", *Expert Systems with Applications*, Vol. 38, No. 3, pp. 1835-1838, 2011.

[7] A.N. Benaichouche, H. Oulhadj and P. Siarry, "Improved Spatial Fuzzy C-means Clustering for Image Segmentation using PSO Initialization, Mahalanobis Distance and Post-Segmentation Correction", *Digital Signal Processing*, Vol. 23, No. 5, pp. 1390-1400, 2013.

[8] Z. Izakian, M.S. Mesgari and A. Abraham, "Automated Clustering of Trajectory Data using a Particle Swarm Optimization", *Computers, Environment and Urban Systems*, Vol. 55, pp. 55-65, 2016.

[9] J.L. Salmeron, S.A. Rahimi, A.M. Navali and A. Sadeghpour, "Medical diagnosis of Rheumatoid Arthritis using data driven PSO-FCM with Scarce Datasets", *Neurocomputing*, Vol. 232, pp. 104-112, 2017.

[10] A. Saxena *et al.*, "A Review of Clustering Techniques and Developments", *Neurocomputing*, Vol. 267, pp. 664-681, 2017.

[11] D. Hein, A. Hentschel, T. Runkler and S. Udluft, "Particle Swarm Optimization for Generating Interpretable Fuzzy Reinforcement Learning Policies", *Engineering Applications of Artificial Intelligence*, Vol. 65, pp. 87-98, 2017.

[12] J. Valente De Oliveira, A. Szabo and L.N. De Castro, "Particle Swarm Clustering in Clustering Ensembles", *Applied Soft Computing*, Vol. 55, pp. 141-153, 2017.

[13] Marco S. Nobile et al., "Fuzzy Self-Tuning PSO: A Settings-Free Algorithm for Global Optimization", *Swarm and Evolutionary Computation*, 2017.

[14] T.M. Silva Filho, B.A. Pimentel, R.M. Souza and A.L. Oliveira, "Hybrid methods for Fuzzy Clustering based on Fuzzy C-means and Improved Particle Swarm Optimization", *Expert Systems with Applications*, Vol. 42, No. 17, pp. 6315-6328, 2015.

[15] A. Mekhmoukh, and K. Mokrani, "Improved Fuzzy C-Means based Particle Swarm Optimization (PSO) Initialization and Outlier Rejection with Level Set methods for MR Brain Image Segmentation", *Computer Methods and Programs in Biomedicine*, Vol. 122, No. 2, pp. 266-281, 2015.

[16] T. Chaira, "A Novel Intuitionistic Fuzzy C means Clustering Algorithm and its Application to Medical Images", *Applied Soft Computing*, Vol. 11, No. 2, pp. 1711-1717, 2011.

[17] S. Shanthi and V.M. Bhaskaran, "Intuitionistic Fuzzy C-means and Decision Tree Approach for Breast Cancer Detection and Classification", *European Journal of Scientific Research*, Vol. 66, No. 3, pp. 345-351, 2011.

[18] T. Chaira and S. Anand. "A Novel Intuitionistic Fuzzy Approach for Tumour/Hemorrhage Detection in Medical Images", *Journal of Scientific and Industrial Research*, Vol. 70, No. 6, pp. 427-434, 2011.

[19] 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.

[20] P. Kaur, A.K. Soni and A. Gosain, "Robust Intuitionistic Fuzzy C-means Clustering for Linearly and Nonlinearly Separable Data", *Proceedings of IEEE International Conference on Image Information Processing*, pp. 1-6, 2011.

[21] R. Bhargava et al., "Rough Intuitionistic Fuzzy C-means Algorithm and A Comparative Analysis", *Proceedings of 6th ACM India Computing Convention*, pp. 1-23, 2013.

[22] P. Balasubramaniam, and V.P. Ananthi, "Segmentation of Nutrient Deficiency in Incomplete Crop Images using Intuitionistic Fuzzy C-means Clustering Algorithm", *Nonlinear Dynamics*, Vol. 83, No. 1-2, pp. 849-866, 2016.

[23] V.P. Ananthi, P. Balasubramaniam, and C.P. Lim, "Segmentation of Gray Scale Image based on Intuitionistic Fuzzy sets Constructed from Several Membership Functions", *Pattern Recognition*, Vol. 47, No. 12, pp. 3870-3880, 2014.

[24] S. Parvathavarthini, N. Karthikeyani, S. Shanthi, and J M Mohan, "Cuckoo-Search based Intuitionistic Fuzzy Clustering Algorithm", *Asian Journal of Research in Social Sciences and Humanities*, Vol. 7, No. 2, pp. 289-299, 2017.

[25] S. Parvathavarthini, N. Karthikeyani, S. Shanthi, and K. Lakshmi, "Crow-Search-Based Intuitionistic Fuzzy C-Means Clustering Algorithm", *Developments and Trends in Intelligent Technologies and Smart Systems*, pp. 1- 22, 2017.

[26] N.K. Visalakshi, S. Parvathavarthini and K. Thangavel, "An Intuitionistic Fuzzy Approach to Fuzzy Clustering of Numerical Dataset", *Proceedings of International Conference on Computational Intelligence, Cyber Security and Computational Models*, pp. 79-87, 2014.

[27] L.A. Zadeh, "Fuzzy Sets", *Information and control*, Vol. 8, No. 3, pp. 338-353, 1965.

[28] K.T. Atanassov, "Intuitionistic Fuzzy Sets: Past, Present and Future", *Proceedings of 3rd Conference of the European Society for Fuzzy Logic and Technology*, pp. 12-19, 2003.

[29] I.K. Vlachos and G.D. Sergiadis, "The Role of Entropy in Intuitionistic Fuzzy Contrast Enhancement", *Proceedings of International Fuzzy Systems Association World Congress*, pp. 104-113, 2007.

[30] Russell C. Eberhart, Yuhui Shi and James Kennedy, "*Swarm Intelligence*", 1st Edition, Morgan Kaufmann, 2001.

[31] A. Asuncion and D.J. Newman, "UCI Repository of Machine Learning Databases", Ph.D. Dissertation, University of California, 2007.

[32] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Cluster Validity Methods: part I", *ACM SIGMOD Record*, Vol. 31, No. 2, pp. 40-45, 2002.

[33] J. C. Dunn, "Well-Separated Clusters and Optimal Fuzzy Partitions", *Journal of Cybernetics*, Vol. 4, No. 1, pp. 95-104, 1974.

[34] C.J. Van Rijsbergen, "Information Retrieval", Ph.D. Dissertation, Department of Computer Science, University of Glasgow, 1979.