International Journal of Computer Theory and Engineering, Vol. 2, No. 2 April, 2010
1793-8201

# An Intuitionistic Fuzzy Approach to Distributed Fuzzy Clustering

N. Karthikeyani Visalakshi, K. Thangavel, and R. Parvathi

***Abstract*—Due to explosion in the number of autonomous data sources, there is an emergent need for effective approaches to distributed clustering. Intuitionistic Fuzzy Set is a suitable tool to cope with imperfectly defined facts and data, as well as with imprecise knowledge. This paper introduces a novel intuitionistic fuzzy based distributed clustering algorithm, to cluster distributed datasets, without necessarily downloading all the data into a single site. The process is carried out in two different levels: local level and global level. In local level, numerical datasets are converted into intuitionistic fuzzy data and they are clustered independently from each other using modified fuzzy C-Means algorithm. In global level, global centroid is computed by clustering all local cluster centroids. The global centroid is again transmitted to local sites to update the local cluster model. The new algorithm is compared against two existing ensemble based distributed clustering algorithms and centralized clustering where all the data are merged into a single data source and clustered. The simulated experiments described in this paper confirm good performance of the proposed algorithm.**

***Index Terms*—Distributed Clustering, Fuzzy C-Means, Local Centroid, Global Centroid, Intuitionistic Fuzzy Sets.**

## I. Introduction

Clustering is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the others. Clustering techniques are applied in many application areas such as data analysis, pattern recognition, image processing, and information retrieval [1]. Today's large-scale datasets are usually logically and physically distributed, requiring a distributed approach to clustering. Huge amounts of data are stored in autonomous, geographically distributed sources over networks with limited bandwidth and large number of computational resources [2].

Traditional Centralized Clustering (CC) methods require all data to be located at the place, where they are analyzed and cannot be applied in the case of multiple distributed datasets, unless all data are transferred to a single location and clustered. Due to technical, economical or security reasons, it is not always possible to transmit all data from different local sites to single location and then perform global clustering. It is obvious that alternate distributed clustering algorithms [3], [4] reduce the communication overhead, central storage requirements and computation times by exchanging few data and avoiding synchronization as much as possible.

Most of the existing distributed clustering algorithms available in the literature [3]-[11] aim to provide hard clusters [12] using K-Means algorithm or density based algorithm. Conventional hard clustering methods restrict each object of the dataset to exactly one cluster. Fuzzy clustering generates a fuzzy partition based on the idea of partial membership expressed by the degree of membership of each object in a given cluster and Fuzzy C-Means (FCM) [12] is one of the most common fuzzy clustering techniques. However, there are very few distributed fuzzy clustering algorithms [13]-[16] which incorporate FCM, to generate distributed fuzzy clusters.

In the real world, distributed clustering applications frequently involve disparate datasets that also consist of inconsistencies or outliers, where it is difficult to obtain homogeneous and meaningful global clusters. Intuitionistic Fuzzy Sets (IFS) [17] are generalized fuzzy sets that are useful in coping with the hesitancy originating from imperfect and imprecise information. Recently, limited attention has been paid in proposing intuitionistic fuzzy based clustering for centralized environment [18]-[21]. But, it is proved that intuitionistic fuzzy based FCM clustering can be more efficient and more effective than the well established FCM algorithm. Hence, it is required to incorporate intuitionistic fuzzy approach with distributed fuzzy clustering to deal with uncertainty among the dispersed data objects and obtain effective and efficient fuzzy clusters in distributed environment. This paper proposes a novel Intuitionistic Fuzzy based Distributed Fuzzy Clustering (IFDFC) algorithm and confirms its superior performance by comparing with two recent distributed clustering algorithms, Distributed K-Means (DKM) and Improved Distributed Combining (IDC) as well as equal performance with CC.

The rest of this paper is organized as follows: Section 2 discusses the related works. Section 3 presents fuzzy clustering of IF data. Section 4 describes distributed fuzzy clustering of IF data. Section 5 summarizes the experimental analysis performed with benchmark datasets. Finally, Section 6 concludes the paper.

Manuscript received October 9, 2009. N. Karthikeyani Visalakshi is with the Department of Computer Science, Vellalar College for Women, Erode – 638 012, Tamilnadu, India (phone: 91-424-2431936, e-mail: karthichitru@yahoo.co.in).

K. Thangavel is with the Department of Computer Science, Periyar University, Salem – 636 011, Tamilnadu, India (e-mail: drktvelu@yahoo.com).

R.Parvathi is with the Department of Mathematics, Vellalar College for Women, Erode – 638 012, Tamilnadu, India (e-mail: paarvathis@rediffmail.com).

## II. Related Works

There are various distributed clustering solutions proposed in the literature and their comprehensive survey can be obtained from [3],[5]. This section reviews the recent research works on distributed clustering and intuitionistic fuzzy based centralized clustering.

The P2P K-Means algorithm is proposed in [3] for distributed clustering of data streams in a peer-to-peer sensor network environment. Jin R. et al. [6] presented distributed version of Fast and Exact K-Means (FEKM) algorithm, which collected sample data from each data source, and communicated it to the central node. The main data structure of FEKM, the cluster abstract table is computed and sent to all data sources to get global clusters.

Jongil Jeong et al. [5] proposed a distributed clustering scenario and modified K-Means algorithm for clustering huge quantities of biological data. By this algorithm, after clustering local datasets using K-Means, local centroids are transferred to central site and average centroid is calculated. Then local datasets are again clustered using averaged centroid as initial central set in K-Means algorithm.

Lamine M. Aouad et al. [7] proposed a lightweight distributed clustering technique based on a merging of independent local sub clusters according to an increasing variance constraint. The key idea of this algorithm is to choose a relatively high number of clusters locally, or an optimal local number using an approximation technique, and to merge them at the global level according to an increasing variance criterion which requires a very limited communication overhead.

Cormode G. et al. [8] have introduced the problem of continuous, distributed clustering, and given a selection of algorithms, based on the paradigms of local vs. global computations, and furthest point or parallel guessing clustering. In their experimental evaluation, the combination of local and parallel guessing addressed the least communication cost.

In [9], Zhou A. et al. proposed an Expectation Maximization based framework to effectively cluster the distributed data streams. In the presence of noisy or incomplete data records, their algorithms learn the distribution of underlying data streams by maximizing the likelihood of the data clusters.

Le-Khac N. et al. [10] presented an approach for distributed density-based clustering. The local models are created by DBSCAN at each node of the system and these local models are aggregated by using tree based topologies to construct global models. In [2], P-SPARROW algorithm is proposed for distributed clustering of data in peer-to-peer environments. The algorithm combined a smart exploratory strategy based on a flock of birds with a density-based strategy to discover clusters of arbitrary shape and size in spatial data.

G. Ji and X. Ling [11] derived the distributed clustering model through ensemble learning and proposed Distributed K-Means Means (DKM). This algorithm first performs local clustering using K-Means, and then sends all mean values to central site; finally global mean values of underlying global clustering are obtained by using K-Means again. None of the techniques discussed above were developed to fuzzy membership cluster results in distributed environment. There are very few distributed clustering algorithms that incorporate fuzzy C-Means technique in the literature [13]-[16]

In [13] Prodip Hore and L. Hall proposed Distributed Combining Algorithm (DCA) to cluster large scale datasets without clustering all the data at a time. Data is randomly divided into almost equal size disjoint subsets. Each subset is clustered using the hard K-Means or fuzzy C-Means algorithm. The centroids of subsets form an ensemble. A centroid correspondence algorithm transitively solves the correspondence problem among the ensemble of centroids. When the number of clusters in each subset is large, the complexity increases in centroid mapping due to collision. Moreover, when the number of clusters in each dataset is different, this type of centroid mapping is found not suitable.

Though fuzzy C-Means is used for local clustering process, the global clustering process produces only hard clusters by comparing the Euclidean distance between the object and global centroid. P. Hore extended this algorithm in [14], [15], to avoid collision and filter bad centroids, but limited to same number of clusters in each data source. This algorithm is also restricted to hard clusters.

The Improved Distributed Combining Algorithm (IDCA) [22] is a refined version of Distributed Combining Algorithm [13], designed for distributed hard clustering. The process of centroid mapping is performed effectively, with the support of Hungarian method of unbalanced assignment problem, when each dataset produces different number of clusters.

R. Kashef and M. S. Kamel [16] proposed a Distributed Cooperative Hard-Fuzzy Clustering (DCHFC) model for document clustering. This model is based on the intermediate cooperation between the hard distributed K-Means and fuzzy distributed C-Means to enhance the performance of the K-Means reduce the computational time taken by the fuzzy algorithm and produce a better global solution.

In [18], Vicenc Torra et al. introduced a method to define intuitionistic fuzzy partitions from the result of different fuzzy clustering algorithms such as FCM, entropy based FCM and FCM with tolerance. In this approach, the intuitionistic fuzzy partition permits to cope with the uncertainty present in the execution of different fuzzy clustering algorithms with the same data and with the same parameterization.

In [19], Z. Xu. et al. have developed a straightforward and practical algorithm for clustering IFS, which consists of the following two steps: Firstly, it employs the derived association coefficients of IFS to construct an association matrix, and utilizes a procedure to transform it into an equivalent association matrix. Secondly, it constructs the $\alpha$-cutting matrix of the equivalent association matrix, and then classifies the IFS under the given confidence levels.

In [20], [21], N. Pelekis et al. investigated the issue of clustering intuitionistic fuzzy representation of images. For this, they proposed a clustering approach based on the FCM algorithm utilizing a novel similarity metric defined over IFS. The performance of the modified FCM algorithm is evaluated for object clustering in the presence of noise and image segmentation. It is proved that clustering intuitionistic fuzzy image representations is more effective, noise tolerant

and efficient as compared with the conventional FCM clustering of both crisp and fuzzy image representations.

## III. Fuzzy Clustering Of Intuitionistic Fuzzy Data Objects

### A. Intuitionistic Fuzzy Sets

Fuzzy sets [23] are many-valued logic which determines only the degree of membership. But, Krassimir T. Atanassov [16] introduced a new component which also determines the degree of non-membership in defining intuitionistic fuzzy set theory. An intuitionistic fuzzy set is defined as a generalization of a fuzzy set.

**Definition 3.1**. Let a set E be fixed. A fuzzy set on E is an object $\overline{A}$ of the form

$$\overline{A} = \left\{ \left\langle x, \mu_{\overline{A}}(x) \right\rangle \middle| x \in E \right\} \quad (1)$$

where $\mu_{\overline{A}} : E \to [0,1]$ defines the degree of membership of the element $x \in E$ to the set $\overline{A} \subset E$. For every element $x \in E$, $0 \le \mu_{\overline{A}}(x) \le 1$.

**Definition 3.2.** An IFS A is an object of the form

$$A = \left\{ \left\langle x, \mu_A(x), \nu_A(x) \right\rangle \middle| x \in E \right\} \quad (2)$$

where $\mu_A : E \to [0,1]$ and $\nu_A : E \to [0,1]$ define the degree of membership and non-membership, respectively, of the element $x \in E$ to the set $A \subset E$. For every element $x \in E$, it holds that $0 \le \mu_A(x) + \nu_A(x) \le 1$.

For every $x \in E$, if $\nu_A(x) = 1 - \mu_A(x)$, then A represents a fuzzy set. The function

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \quad (3)$$

represents the degree of hesitancy of the element $x \in E$ to the set $A \subset E$.

### B. Modified Fuzzy C-Means Clustering

In [20], Pelekis proposed a new variant of FCM clustering algorithm that copes with uncertainty and a similarity measure between intuitionistic fuzzy sets, based on the membership and non membership values of their elements. The procedure used in modified FCM is same as conventional FCM, except in similarity measure used to compute the membership degree of the object to cluster. Instead of Euclidean distance in conventional FCM, the modified FCM applies IF similarity measure for any two elements namely *A* and *B* as follows:

$$S_1(A,B) = \frac{S'(\mu_A(x_i), \mu_B(x_i)) + S'(\nu_A(x_i), \nu_B(x_i))}{2} \quad (4)$$

where

$$S'(A',B') = \begin{cases} \dfrac{\sum_{i=1}^{n} \min(A'(x_i), B'(x_i))}{\sum_{i=1}^{n} \max(A'(x_i), B'(x_i))}, & A' \cup B' \neq \Phi \\ 1, & A' \cup B' = \Phi \end{cases} \quad (5)$$

The modified FCM algorithm is described in Fig. 1. Initially, *C* numbers of centroids are randomly selected from the intuitionistic fuzzy objects, which contain both membership and non-membership values. Next, the membership degree of each object to each cluster $U_{ij}$ is computed using IF similarity measure as in equation (6). The centroids are then updated using Cluster Membership Matrix (CMM) $U_{ij}$ and corresponding membership and non-membership degrees of centroids $V_i$ are also computed. The above two steps are repeated, until it reaches convergence.

**Algorithm**. Modified FCM

**Input** : Dataset of *n* objects with *d* features, value of *C* and fuzzification value *m* > 1

**Output:** Cluster Membership Matrix $U_{ij}$ for n objects and *C* clusters

**Procedure:**

*Step 1:* Determine initial centroids by selecting *c* random intuitionistic fuzzy objects.

*Step 2:* Compute the values for CMM represented by $U_{ij}$, using

$$\underset{\substack{1 \le i \le c \\ 1 \le j \le n}}{\forall} U_{ij} = \begin{cases} \dfrac{\left(S_1(x_j - V_i)\right)^{\frac{1}{1-m}}}{\sum_{l=1}^{c} \left(S_1(x_j - V_l)\right)^{\frac{1}{1-m}}}, & I_j = \phi \\ \begin{cases} 0, & i \notin I_j \\ \sum_{i \in I_j} U_{ij} = 1, & i \in I_j, I_j \neq \phi \end{cases} \end{cases} \quad (6)$$

where

$$\underset{\forall 1 \le j \le n}{I_j} = \left\{ i \middle| 1 \le i \le c; S_1(x_j, V_i) = 0 \right\}$$

*Step 3:* Update the centroids' matrix $V_i$ using

$$\underset{1 \le i \le c}{\forall} V_i = \frac{\sum_{j=1}^{n} (U_{ij})^m x_j}{\sum_{j=1}^{n} (U_{ij})^m} \quad (7)$$

*Step 4:* Compute membership and non-membership degrees of $V_i$

*Step 5:* Repeat step 2 to step 4 until converges.

Fig. 1. Modified Fuzzy C-Means Algorithm

### C. Intuitionistic Fuzzy Representation of Numerical Data Objects

The proposed IFDFC algorithm requires that each element is to be converted into a pair of membership and non membership values. A new procedure for intuitionistic fuzzy representation of numeric data is derived, by modifying the definition for intuitionistic fuzzy representation of digital image [24]. In this process, the crisp dataset is first

transferred to fuzzy domain and sequentially into the intuitionistic fuzzy domain, where the clustering is performed.

Let *X* be the dataset of *n* objects and each object contains *d* features. The proposed IF data clustering requires that each data element $x_{ij}$ belongs to IFS $X'$ by a degree $\mu_i(x_j)$ and does not belong to $X'$ by a degree $\nu_i(x_j)$, where *i* and *j* represent objects and features of the dataset respectively.

A membership function $\overline{\mu_i}(x_j)$ for intermediate fuzzy representation is defined by

$$\overline{\mu_i}(x_j)=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \tag{8}$$

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

The intuitionistic fuzzification based on the family of parametric membership and non-membership function, used for clustering, is defined respectively by

$$\mu_i(x_j;\lambda)=1-(1-\overline{\mu_i}(x_j))^{\lambda} \tag{9}$$

and

$$\nu_i(x_j;\lambda)=(1-\overline{\mu_i}(x_j))^{\lambda(\lambda+1)} \tag{10}$$

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

The intuitionistic fuzzification converts crisp dataset $X(x_{ij})$ into intuitionistic fuzzy dataset $X'(x_{ij},\mu_i(x_j),\nu_i(x_j))$.

## IV. Distributed Clustering of Intuitionistic Fuzzy Data Objects

### A. Distributed Clustering

The main objective of distributed clustering algorithms is to cluster the distributed datasets without necessarily downloading all the data to the single site. It assumes that the objects to be clustered reside on different sites. This process is carried out in two different levels: local level and global level. In local level, all sites carry out clustering process independently from each other. After having completed the clustering, a local model such as cluster centroids is determined, which should reflect an optimum trade-off between complexity and accuracy. Next, the local model is transferred to a central site, where the local models are merged in order to form a global model. The resultant global model is again transmitted to local sites to update the local models [25].

The key idea of distributed clustering is to achieve a global clustering that is as good as the best centralized clustering algorithm with limited communication required to collect the local models or local representatives into a single location, regardless of the crucial choice of any clustering technique in local site**.** Distributed clustering algorithms [3] can be classified along two independent dimensions such as classification based on data distribution and data communication.

A common classification based on data distribution in the literature [3], [4] is those, which apply to homogeneously distributed or heterogeneously distributed data. Homogeneous datasets contain the same set of attributes across distributed data sites. Examples include local weather databases at different geographical locations and market-basket data collected at different locations of a grocery chain. Heterogeneous data model supports different data sites with different schemata. For example, a disease emergence detection problem may require collective information from a disease database, a demographic database and biological surveillance databases.

According to the type of data communication, distributed clustering algorithms are classified into two categories: multiple communications round algorithms and centralized ensemble-based algorithms. The first group consists of methods requiring multiple rounds of message passing. These methods require a significant amount of synchronization, whereas the second group works in an asynchronous manner, first generating the local clusters and then combining those at the central site [26].

### B. Proposed Algorithm

The step by step procedure of proposed Intuitionistic Fuzzy based Distributed Fuzzy Clustering algorithm for homogeneously distributed datasets is described in Fig. 2. First, minimum and maximum values of each feature vectors are extracted from all local datasets and transmitted to central place, where global minimum and maximum values are identified. These two values are used to convert real scalar values of local datasets into pair of global IF data objects using the equations (9) and (10). Next, the IF objects of local datasets are clustered using modified FCM to obtain CMM and local centroids in terms of membership and non membership. All local centroids are merged into a pair of centroids datasets and clustered using the same modified FCM algorithm at central place, to group similar centroids and obtain global centroids. The CMM of local datasets are then updated using global centroids to obtain global fuzzy clusters of distributed datasets.

**Algorithm.** IFDFC
**Input :** Homogeneous *p* datasets, each with *d* dimensions
**Output:** Global fuzzy clusters of *p* datasets

**Procedure:**
*Step 1:* Find maximum and minimum values of each feature from each local dataset and transmit them into central place
*Step 2*: Compute global maximum and minimum value at central place
*Step 3:* Convert real scalar values of local datasets into IF values using equation (9) and equation (10)
*Step 4:* Cluster each local IF dataset by modified FCM algorithm and obtain CMM and IF form of cluster centroids
*Step 5:* Merge membership and non membership values of cluster centroids of local datasets into a pair of centroids datasets

*Step 6:* Cluster centroids datasets using modified FCM to obtain global centroids
*Step 7:* Update CMM using global centroid to obtain global fuzzy clusters

Fig. 2 Intuitionistic Fuzzy based Distributed Fuzzy Clustering

## V. Experimental Analysis

In this section, empirical evidence is provided for distributed fuzzy clustering, that the high quality global cluster models is obtained with limited communication overhead and high level of privacy. The efficiency of IFDFC is compared against two recent distributed clustering algorithms, DKM and IDC along with CC, where all local datasets are merged, converted into IF data and clustered using modified FCM. All experiments are conducted with the assumption of having non-overlapping objects with same set of features in distributed datasets, for both uniform and non-uniform type of data distribution.

### A. Experimental Setup

The algorithms have been implemented and tested with six bench mark numeric datasets available in the UCI machine learning data repository [27]. The information about the datasets is shown in Table I. For the purpose of experimental setup, the dataset is divided into different disjoint subsets and each subset is considered as distributed data source. The experiment on each dataset runs 25 times and the average results are considered for analysis. The value of $\lambda$ is set as 0.95, for the computation of membership and non-membership values, irrespective of the characteristics of the datasets.

### B. Evaluation Methodology

The performance of the proposed algorithm is measured in terms of three external validity measures [28], [29] namely Rand index, F-Measure and Entropy. The external validity measures test the quality of clusters by comparing the results of clustering with the 'ground truth' (true class labels). The Rand index measures the agreement between true class labels and cluster results. The F-Measure measures the extent to which a cluster contains only objects of a particular class and all objects of the class. The Entropy is used to measure the degree to which each cluster consists of objects of a single class. In case of Rand index and F-Measure, the value 1 indicates that the data clusters are exactly same and so the increase in the values of these measures proves the better performance. But, the value 0 signifies that the data clusters are perfect for Entropy measure and so the value of this measure is to be decreased to reach better quality clusters.

### C. Uniform Type of Data Distribution

In uniform type of data distribution, the cardinality of each subset has been kept as almost equal and number of clusters produced by each subset has also been kept as equal. In first experiment, all datasets are divided into three subsets, under uniform type of data distribution and the algorithms are evaluated. The results of IFDFC, in comparison with the results of DKM, IDC and CC, in terms of Rand index, F-Measure and Entropy are shown in Table II, Table III and Table IV respectively. From the Tables, it is observed that IFDFC algorithm yields better results than DKM and IDC algorithms for all datasets, in terms of Rand index and Entropy. According to F-Measure, the performance of IFDFC algorithm dominates the performance of DKM and IDC algorithms except for Australian dataset. The values of both F-Measure and Entropy are highly appreciable for dermatology dataset with IFDFC algorithm. It is noted that the quality of clusters produced by IFDFC is as good as CC roughly for all datasets. The average performance of these three algorithms in terms of Rand Index, F-Measure and Entropy are depicted in Fig. 3, Fig. 4 and Fig. 5 respectively.

In next experiment, the segmentation dataset is divided into different number of subsets, in order to evaluate the scalability of the proposed distributed clustering algorithm. Table V shows the performance of IFDFC on segmentation dataset, for different number of subsets. First row of the table (Segment -1S) depicts the results of Rand index, F-Measure and Entropy, when all objects are kept in a single place, second row (Segment – 3S) shows the same type of results, when the objects are divided into 3 subsets, and so on. Here, it is proved that the proposed algorithm is consistent, independent of the number of subsets. It is also noted that the value of entropy decreases, when the number of subsets are increased.

TABLE I. Details of Datasets

| S. No. | Dataset | No. of Attributes | No. of Classes | No. of Instances |
|---|---|---|---|---|
| 1 | Australian | 14 | 2 | 690 |
| 2 | Breast Cancer | 10 | 2 | 699 |
| 3 | Dermatology | 34 | 6 | 366 |
| 4 | Mammography | 5 | 2 | 961 |
| 5 | Pen Digit | 16 | 10 | 10992 |
| 6 | Segmentation | 19 | 7 | 2310 |

TABLE II. Comparative Analysis Based on Rand Index

| S. No. | Dataset | DKM | IDC | CC | IFDFC |
|---|---|---|---|---|---|
| 1 | Australian | 0.511 | 0.512 | 0.750 | 0.746 |
| 2 | Breast Cancer | 0.521 | 0.523 | 0.944 | 0.943 |
| 3 | Dermatology | 0.682 | 0.699 | 0.905 | 0.855 |
| 4 | Mammography | 0.561 | 0.558 | 0.657 | 0.650 |
| 5 | Pen Digit | 0.845 | 0.892 | 0.925 | 0.913 |
| 6 | Segmentation | 0.795 | 0.803 | 0.865 | 0.863 |

TABLE III. Comparative Analysis Based on F-Measure

| S. No. | Dataset | DKM | IDC | CC | IFDFC |
|---|---|---|---|---|---|
| 1 | Australian | 0.585 | 0.576 | 0.479 | 0.488 |
| 2 | Breast Cancer | 0.630 | 0.634 | 0.971 | 0.974 |

| | | | | | |
|---|---|---|---|---|---|
| 3 | Dermatology | 0.290 | 0.288 | 0.823 | 0.771 |
| 4 | Mammography | 0.668 | 0.671 | 0.779 | 0.779 |
| 5 | Pen Digit | 0.582 | 0.649 | 0.691 | 0.673 |
| 6 | Segmentation | 0.482 | 0.503 | 0.647 | 0.644 |

TABLE IV. COMPARATIVE ANALYSIS BASED ON ENTROPY

| S. No. | Dataset | DKM | IDC | CC | IFDFC |
|---|---|---|---|---|---|
| 1 | Australian | 0.356 | 0.338 | 0.188 | 0.186 |
| 2 | Breast Cancer | 0.276 | 0.254 | 0.066 | 0.071 |
| 3 | Dermatology | 0.987 | 0.947 | 0.193 | 0.205 |
| 4 | Mammograph | 0.611 | 0.619 | 0.489 | 0.490 |
| 5 | Pen Digit | 0.651 | 0.670 | 0.619 | 0.621 |
| 6 | Segmentation | 0.939 | 0.965 | 0.601 | 0.536 |

TABLE V. COMPARATIVE ANALYSIS FOR DIFFERENT NUMBER OF SUBSETS

| S. No. | Dataset | Rand Index | F-Measure | Entropy |
|---|---|---|---|---|
| 1 | Segment -1S | 0.863 | 0.647 | 0.601 |
| 2 | Segment -3S | 0.858 | 0.644 | 0.536 |
| 3 | Segment -5S | 0.862 | 0.654 | 0.528 |
| 4 | Segment -7S | 0.859 | 0.641 | 0.511 |
| 5 | Segment -10S | 0.855 | 0.639 | 0.510 |

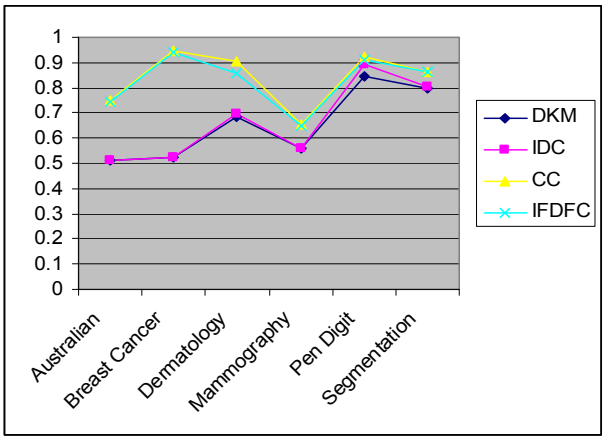


Fig. 3. Comparative Analysis based on Rand Index

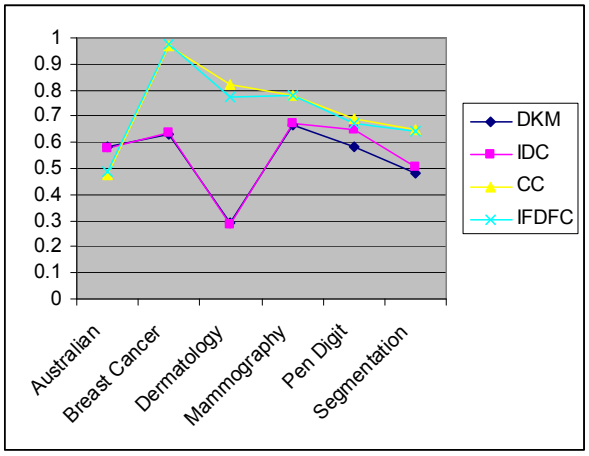


Fig. 4. Comparative Analysis based on F-measure

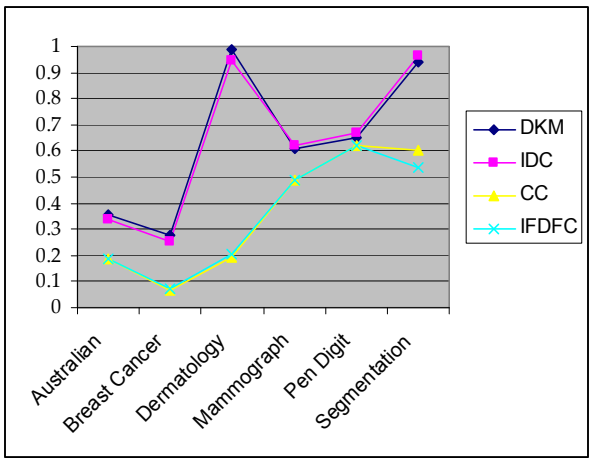


Fig. 5. Comparative Analysis based on Entropy

### *D. Non-Uniform Type of Data Distribution*

In non-uniform type of data distribution, it is assumed that individual data sources are having varying numbers of clusters and varying types of clusters. In some applications, one or more classes may be missing in some data sources and some times data sources may have entirely different types of clusters. For example, data marts of individual stores of a retail company may deal with different types of customers or products, according to the demands in that place. The similar situation is also efficiently handled in the proposed algorithm, by simply providing required number of global clusters, independent of the number of local clusters.

The results of Segmentation dataset in terms of Rand Index, F-Measure and Entropy obtained for this assumption is represented in Table VI, along with the number of objects in three data sources namely S1, S2 and S3. The numbers provided in braces indicate the cluster labels in corresponding data sources.

Though IFDFC provides equal performance as CC, in terms of Rand index, F-Measure and Entropy, it outperforms CC in terms of communication overhead, space complexity and privacy maintenance. In CC, all objects are to be transferred to central place and fuzzy clustering algorithm is to be executed to find global membership value. In real application scenario, it needs huge communication cost, since many data sources may contain large number of high dimensional data objects. In distributed approach, only centroids of local clusters and global centroid is to be

transmitted between data sources. The centroids are insensitive to a number of objects in each data source and the size of cluster centroid is definitely much less than the size of data objects or even the size of label vector. Moreover, centralized clustering needs much memory space at one place, according to the size of objects accumulated for clustering. It remains important to preserve privacy on individual objects for most of the distributed nature of application scenario like financial, banking and medical applications. Since centroids of clusters represent only prototype, the proposed method of distributed clustering enables privacy preserving data mining framework.

TABLE VI. RESULTS OF NON-UNIFORMLY DISTRIBUTED SEGMENTATIONDATASET (NUMBER OF GLOBAL CLUSTERS – 7)

| S. No. | Size of datasets (cluster labels) | | | Rand Index | F-Measure | Entropy |
|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | | | |
| 1 | 700 (1, 2, 3. 4. 5) | 770 (1, 2, 3, 6, 7) | 840 (1, 2, 3, 4, 6, 7) | 0.863 | 0.647 | 0.601 |
| 2 | 750 (2, 4, 6, 7) | 752 (1, 2, 3, 5, 6) | 808 (1, 2, 3, 4, 5) | 0.848 | 0.654 | 0.612 |
| 3 | 762 (1, 2, 3, 4, 7) | 843 (2, 3, 5, 6, 7) | 705 (1, 2, 3, 5 ,6) | 0.834 | 0.654 | 0.603 |
| 4 | 554 (1, 2, 3, 4) | 642 (2, 3, 5, 6, 7) | 1114 (1, 2, 3, 4, 5, 6, 7) | 0.862 | 0.651 | 0.601 |

## VI. CONCLUSION

A novel method of distributed fuzzy clustering using intuitionistic fuzzy set theory is proposed to produce fuzzy clusters in distributed environment. Comprehensive experiments on six benchmark numerical datasets have been conducted to study the impact of using intuitionistic fuzzy approach in distributed fuzzy clustering. It can be concluded that the proposed algorithm leads to obtain better quality clusters than the two existing distributed clustering algorithms. At the same time, it is also proved that the performance of the proposed algorithm is almost same as the performance of centralized clustering. In future, optimization algorithm will be applied for tuning of parameter $\lambda$ to produce superior quality clusters. The on going research also focuses, in particular, on enhancing proposed clustering algorithm to produce intuitionistic fuzzy partitions, in centralized as well as distributed environment.

## REFERENCES

[1] Jain A. K. Murthy M. N. Flynn P. J. Data clustering: A review, ACM Computing Surveys, 31 (3), 1999, pp. 265-323.
[2] Folino G. Forestiero A. Spezzano G. Swarm-based distributed clustering in peer-to-peer systems, In Artificial Evolution, Lecture Notes in Computer Science, Talbi E. et al. (Eds.), Springer-Verlag, 2006, pp. 37-48.
[3] Sanghamitra B. Giannella C. Maulik U. Kargupta H. Liu. K. Datta S. Clustering distributed data streams in peer-to-peer environments, Information Science, 176 (4), 2006, pp. 1952-1985.
[4] Ghosh J. Merugu S. Distributed clustering with limited knowledge sharing, In Proceedings of the 5th International Conference on Advances in Pattern Recognition, Calcutta, India, December 11-13, 2003, pp. 48-53.
[5] Jin R. Goswami A. Agarwal G. Fast and exact out-of-core and distributed K-Means clustering, Knowledge and Information Systems, 10 (1), 2006, pp. 17-40.
[6] Jeong J. Ryu B. Shin D. Shin D, Integration of distributed biological data using modified k-means algorithm, In Emerging Technologies in Knowledge Discovery and Data Mining, LNCS, Washio T. et al. (eds.), Springer-Berlin, 2007, pp. 469-475.
[7] Lamine M. A. Le-Khac N. Tahar M. K. Lightweight clustering technique for distributed data mining applications. In Advances in data mining, Theoretical aspects and applications, LNCS, Perner P. (Ed.), Springer, 2007. pp. 120-134.
[8] Cormode G. Muthukrishnan S. Zhuang W. Conquering the divide: continuous clustering of distributed data streams, In IEEE 23rd International Conference on Data Engineering, Turkey, April 15-20, 2007, pp. 1036-1045.
[9] Zhou A. Cao F. Yan Y. Sha C C. He X, Distributed data stream clustering: a fast EM-based approach, In IEEE 23rd International Conference on Data Engineering, Turkey, April 15-20, 2007, pp. 736-745.
[10] Le-Khac N. Lamine M A. Tahar M K. A new approach for distributed density based clustering on grid platform, In Data Management, Data Data every where, LNCS, Kooper R. Kennedy J. (eds.), Springer-Berlin, 2007, pp. 247-258.
[11] Ji Genlin, Ling Xiaohan, Ensemble learning based distributed clustering, In Emerging Technology and Knowledge Discovery and Data Mining, LNCS, Washio T. et al. (eds.), Springer-Verlag, 2007, pp. 312-321.
[12] Pang-Ning Tan, Steinbach M, Kumar V. Cluster analysis: basic concepts and algorithms. In Introduction to Data Mining, Pearson Addison Wesley, Boston, 2006,
[13] Hore P. Lawrence O. Hall, Scalable clustering: A distributed approach, In IEEE International Conference on Fuzzy Systems, Hungary, July 25-29. 2004, pp. 25-29.
[14] Hore P. Lawrence O. Hall, Dimitry B. Goldgofz, A Cluster ensemble framework for large datasets, In Proceedings of IEEE Conference on Systems, Man Cybernetics B, Taiwan, October 8-11, 2006, Vol. 4. pp. 3342-3347.
[15] Hore P. Lawrence O. Hall, Dimitry B. Goldgof, A scalable framework for cluster ensembles, Pattern Recognition, 42(5):, 2008,pp. 676-688.
[16] Kashef R. Kamel M. Distributed cooperative hard-fuzzy document clustering, In Proceedings of the Annual Scientific Conference of the LORNET Reseach Network, Montreal, November 8-10, 2006, pp. 8-10.
[17] Krassimir T. Atanassov. Intuitionistic fuzzy sets: past, present and future. In Proceedings of the 3rd Conference of the European Society for Fuzzy Logic and Technology. Germany, September 10-12, 2003, pp. 12-19.
[18] Torra V. Miyamoto S. Endo Y. Domingo-Ferrer J. On intuitionistic fuzzy clustering for its application to privacy. In Proceedings of IEEE International Conference on Fuzzy Systems. Hong Kong, China, June 1-6, 2008, pp. 1042-1048.
[19] Zeshui Xu, Jian Chen, Junjie Wu. Clustering algorithm for intuitionistic fuzzy sets. Information Sciences , 2008, 178(19), pp. 3775-3790.
[20] Nikos Pelekis, Dimitrios K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis. Fuzzy clustering of intuitionistic fuzzy data. International Journal of Business Intelligence and Data Mining, 3(1), pp. 45-65.
[21] Dimitrios K. Iakovidis, Nikos Pelekis, Evangelos E. Kotsifakos, Ioannis Kopanakis. Intuitionistic fuzzy clustering with applications in computer vision. In Advanced Concepts for Intelligent Vision Systems. LNCS, Blanc-Talon et al. (eds.), Springer Berlin/ Heidelberg, 2008, pp. 764-774.
[22] Karthikeyani Visalakshi N. Thangavel K. Alagambigai P. Ensemble approach to distributed clustering, In Mathematical and Computational Model, Natarajan et al. (eds.), Narosa Publishing House, New Delhi, 2007, pp. 252-261.
[23] Lotfi A. Zadeh, Fuzzy sets, Information and Control, 8(3), 1965, pp. 338-353.
[24] Ioannis K. Vlachos, George D. Sergiadis, The role of entropy in intuitionistic fuzzy contrast enhancement. In Foundations of fuzzy logic and soft computing, LNCS, Melin P. et al. (eds.), Springer Berlin/ Heidelberg, 2007, pp. 104-113.
[25] Januzaj E. Kriegel Hans P. Pfeifle M. DBDC: Density based distributed clustering, In Advances in databases technology – EDBT 2004, LNCS, E. Bertino, S. Christodoulakis, D. Plexousakis (Eds.), Springer Berlin/ Heidelberg, 2004, pp. 529-530.

[26] Park B. Kargupta H. Distributed data mining, The Hand Book of Data Mining, Nong Ye, (ed.), Lawrence Erlabum Associates, Publishers, Mahwah, Newjersey, 2003, pp. 341-358.
[27] Merz C. J. Murphy P. M. UCI repository of machine learning databases, Irvine, University of California, < http://www.ics.uci.eedu/~mlearn/>, 1998.
[28] Halkidi M. Batistakis Y. Vazirgiannis M, Cluster validity methods: part II, ACM SIGMOD Record, 31(3), 2002, pp. 19-27.
[29] Hui Xiong. Junjie Wu, Jian Chen. K-means clustering versus validation measures: a data distribution perspective. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA, August 20-23, 2006, pp. 779-784.

**N. Karthikeyani Visalakshi :** She received M.C.A degree in 1992 and M. Phil degree in 1999 from Bharathiar University, Coimbatore, Tamil Nadu, India. Currently she is working as a Assistant Professor at Vellalar College for Women, Erode, Tamil Nadu, India and her experience in teaching started from the year 1992. She is doing research in the Mother Therasa University, Kodaikanal, India. Her areas of interests include Distributed Data Mining, Clustering, Rough sets, Fuzzy logic and Intuitionistic Fuzzy sets.

**Dr. K. Thangavel: He** has received Ph. D degree in the area of Optimization Algorithms from the Gandhigram Rural Institute-Deemed University, Gandhigram, Tamilnadu, India in 1999. Currently he is working as Professor and Head in the Department of Computer Science, Periyar University, Salem, Tamilnadu, India. He has published more than 125 research publications in various National and Inter National Journals. He has edited three books published by Narosa Publishers NewDelhi, India. He has successfully guided for 5 Ph. D and 10 M. Phil scholars. More than 15 Ph. D scholars are pursuing under his research supervision. His research interest includes Data Mining, Digital Medical Image Processing, Soft Computing, Mobile Computing and Bio-informatics. He is a life member of Operational Research Society of India and member of the research group in Rough set Society. He has organized three National Conferences, three National Seminars, five Research workshops and two 21 day UGC refresher progrmmes. He is reviewer for leading publishers such as Elseviewer, Springer, Taylor and Francis.

**Dr. R.Parvathi:** she is working as Assistant Professor in Mathematics. Completed Post Graduate in Mathematics (1989), M.Phil (1990) and Ph.D., (2006) in Alagappa University, India. She has got 17 years of teaching experience and 8 years of research experience. Her Research Area is Intuitionistic Fuzzy Sets (IFSs): Theory and Applications. She has developed algorithms in Image Processing, Shortest Paths in Networks and Morphological Operators using IFSs. She has completed UGC Minor Research Project. She was a Summer Research Fellow (2007) of Indian Academy of Sciences, Indian National Science Academy and National Academy of Sciences, India. At present she is working on a Project under UGC Research Award and Indo – Bulgarian Bilateral Research Programme. She has visited UK, Germany and Bulgaria to present papers in International Conferences. She has published more than 15 Research papers in reputed International Journals and she took part in more than 35 conferences at National and International level as paper presenter/ Resource person.