EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

A new Semi-Supervised Intuitionistic Fuzzy C-means Clustering

J. Arora[1,]* and M. Tushir[2]

[1]Assistant Professor, Dept. of Information Technology, MSIT, Affiliated to GGSIPU, Delhi, India
[2] Professor, Dept. of Electrical & Electronics Engineering, MSIT, Affiliated to GGSIPU, Delhi, India

Abstract

Semi-supervised clustering algorithms aim to increase the accuracy of unsupervised clustering process by effectively exploring the limited supervision available in the form of labelled data. Also the intuitionistic fuzzy sets, a generalization of fuzzy sets, have been proven to deal better with the problem of uncertainty present in the data. In this paper, we have proposed to embed the concept of intuitionistic fuzzy set theory with semi-supervised approach to further improve the clustering process. We evaluated the performance of the proposed methodology on several benchmark real data sets based on several internal and external indices. The proposed Semi-Supervised Intuitionistic Fuzzy C-means clustering is compared with several state of the art clustering/classification algorithms. Experimental results show that our proposed algorithm is a better alternative to these competing approaches.

Keywords: Semi-Supervised Clustering, Intuitionistic Fuzzy-Set, Image Segmentation

Received on 24 February 2019, accepted on 10 July 2019, published on 24 July 2019

doi: 10.4108/eai.13-7-2018.159622

1. Introduction

Nowadays, Data mining techniques have been recognized in different fields for discovering useful patterns and extracting information from the pool of available abundant data. Data mining provides automated tools for the process of knowledge discovery by analyzing available data. It helps in measuring the degree of membership and non-membership by analyzing different parameters in the dataset [1]. Based on the information processed, dataset can also be grouped into different classes. At basic level data mining techniques can be broadly categorized as predictive and descriptive method. Fig. 1 shows details of different types of data mining techniques. In predictive methods, classification is most commonly used for the process of grouping the data into different classes. The process of classification is based on the availability of two sets of data i.e. training and testing. The commonly used classification techniques are naive bayes, support vector machines, decision trees, neural networks etc. [2]. Classification is a difficult process, as system requires proper training with respect to the features that have to be extracted. These involve statistical and probabilistic methods for the process of data analytics and data mining. The performance of classification process depends on the availability of knowledge of data to be analyzed. But, the acquisition of data point’s knowledge (labeling) is always a costly, error-prone and tedious process. The process of classification can also be termed as supervised clustering.

Among descriptive methods, clustering is one of the most relevant technique for data mining, pattern matching and discovery of knowledge [3]. The purpose of clustering is to divide the data set into groups, so that similar data points fall into same cluster. Clustering is an unsupervised process, where the data is not supported with labeled information, so its aim is to infer the expected structure existing within a set

[*]Corresponding author. Email:joy.arora@gmail.com

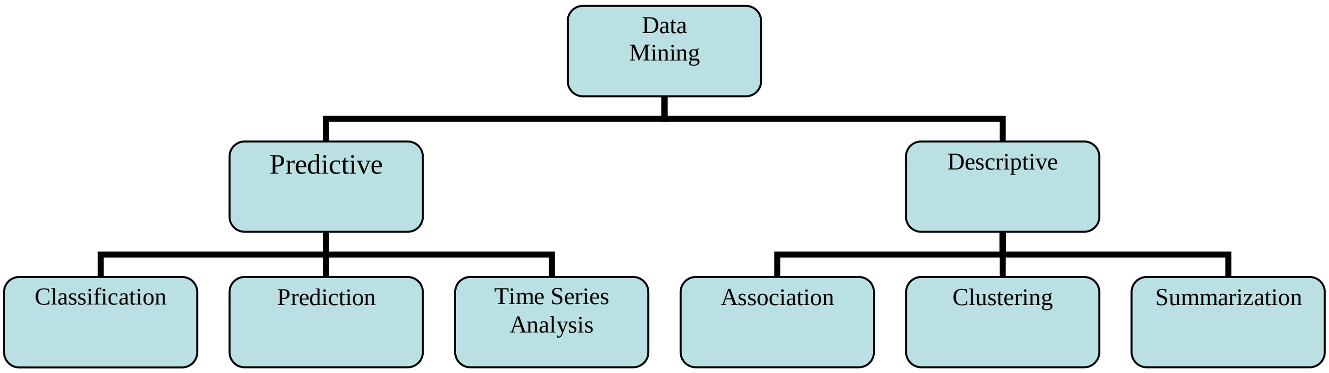

Figure 1. Different types of Data mining techniques

of data points. Secondly, Unsupervised clustering techniques process totally unlabelled data, therefore suffer from the problem of defining number of clusters, prior random initialization of the cluster centers, problem of local traps and finally binding every data point to a class. Nowadays soft computing methods are integrated with probabilistic methods, resulting into more robust and interpretable model which handle better the wide range of information present in the data [4]. To better handle the issues related to supervised and unsupervised clustering, semi-supervised clustering has received a lot of attention in the area of machine learning.

Semi-supervised clustering discovers its application in situations where information is neither completely nor accurately labeled. Semi-supervised clustering deals with the problem of unsupervised clustering, defining more accurate clusters using some labeled data along with unlabelled data, to acquire better clustering results. There are a variety of semi-supervised clustering techniques that have been given and proved to perform better as compared to unsupervised approach. Basu et al. [5] proposed semi-supervised clustering algorithm based on center initialization mechanism. In this algorithm, seeds are used to initialize the centers of clusters using labeled data and then updated using clustering process. Demiriz et al. [6] have used genetic algorithm along with supervised and unsupervised clustering to design semi-supervised clustering algorithm. Blum et al [7] used graph based method to provide information regarding labeling in the process of unsupervised clustering. Dinler et al [8] have given the semi-supervised clustering algorithm that aims to partition regional data objects in the presence of instance level constraints. Saha et al [9] proposed the concept of semi-supervised clustering using multiobjective optimization and applied the concept in the process of automatic medical image segmentation. In this paper, 10% of the pixels are labeled for the initial class values. Despite the promising results [10]-[12], a major challenge often arises in dealing with unlabelled data in semi-supervised clustering, in the presence of noise and uncertainty.

Generally the percentage of unlabelled data is more as compared to labeled data. So it becomes a difficult task to handle the uncertainty of unlabelled data. This problem is widely related to many real world problems [13]-[14]. Recently intuitionistic theory has been widely used with classical clustering algorithms to deal with the problem of uncertainty present in the real world unlabelled data. Different experiments have proved [15]-[17] that intuitionistic fuzzy set based method helps to better handle the problem of uncertainty as compared to fuzzy set. Intuitionistic fuzzy set (IFS) is a higher order extension of fuzzy set. Intuitionistic fuzzy sets are elaborated set, consisting of hesitation degree along with membership and non-membership degree. The hesitation degree helps to deal with the problem of uncertainty present in the unlabelled data. In literature, few researchers have used IFS effectively in different applications. Among these works, Xu and Wu [18] defined IFCM which presents the clustering of Intuitionistic fuzzy set (IFS). Pelekis et al. [19] demonstrated the process of clustering based on the intuitionistic fuzzy knowledge of data and recommended that the intuitionistic fuzzy clustering acquires the qualitative information, which may be estimated as per feature vector. Chaira [20] has given intuitionistic fuzzy clustering algorithm for the process of medical image segmentation which exploits the benefits of IFS. Although IFS based clustering algorithm proved to give better result as comparison to fuzzy set based clustering approaches but still it suffers the same problem as that of unsupervised clustering.

This paper proposes a new semi-supervised clustering technique by embedding the concept of intuitionistic fuzzy set theory to handle uncertainty present in the unlabelled data. In the proposed method, a small set of labeled data is used to provide partial supervision to the unsupervised clustering approach. The partial supervision along with intuitionistic set theory is used to further label the unlabelled data.

The proposed technique will allow overcoming several problems associated with clustering process: (i) defining the number of clusters (ii) the problem of random initialization of cluster centers (iii) handling the uncertainty present in the unlabelled data (iv) sensitivity to noise and outliers.

The rest of this paper is sorted out as follows. Section 2 gives the brief overview of data mining learning techniques.

Section 3 gives the related work on the semi-supervised fuzzy C-means and intuitionistic fuzzy sets. In Section 4, the proposed semi-supervised intuitionistic fuzzy c-means is presented in detail. Section 5 features the capability of the proposed methodology through various experiments on benchmark datasets and on a natural image along with assumptions and limitations. Section 6 presents the concluding remarks.

2. Learning in Data Mining

The concept of data mining has been initiated from three different methods at the intersection i.e. database systems, statistics and artificial intelligence. It helps to analyse the data for the purpose of knowledge discovery (extraction of pattern and knowledge) [10]. Machine learning is a method towards artificial intelligence, which helps the system to learn in different ways. The process of learning starts with the perceptions from the data, including different examples, instructions that can be followed to look for some patterns and helps to make better decisions for future predictions. Fig. 2 gives the broad categorisation of machine learning methods.

In supervised learning methods, an algorithm is learned to map input to output as required. Suppose Y is an output variable and X is an input variable. Here function F is trained to map input to the output variable.

$$Y = F(X) \tag{1}$$

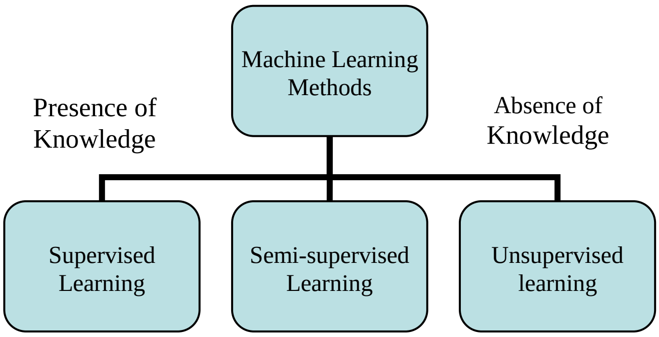

Figure 2. Machine learning methods

The aim of supervised learning process is to estimate the mapping function so that, it can acquire the new input data X and function F can predict the output variable Y for that data. The output achieved can be compared with the correct output, in order to find the errors and train the function F in a better way. The accuracy of learning process highly depends on the availability of the label data, correctness of the data and presence of noise should be minimised [21]. This learning mechanism basically includes classification and regression.

In Unsupervised learning methods, we are provided with only input data and no output data. Here the algorithm is developed to group the available input data in specified number of clusters. The aim of unsupervised learning process is to study the present structure or distribution in the input data in a way to predict more about the data. The process is called unsupervised learning because no knowledge is being provided as supervised learning process. There is no supervision, no training and testing process is carried out to model the algorithm. The algorithms are left to formulate by themselves based on the objective function and presence of internal structure in the data. In this process, number of clusters and their initialization is required to start the process [22]. The common clustering algorithms are k-means, fuzzy c-means, intuitionistic fuzzy c-means etc.

Semi-Supervised learning is an approach that includes benefits of unsupervised learning and supervised learning techniques. In these learning problems, an algorithm is provided with large amount of unlabelled data and some amount of labelled data. A large number of machine learning problems fall in this category. Supervised learning is a tedious, time-consuming and expensive process as it requires large amount of labelled data. Lot of efforts are required by domain experts for the process of labelling the data whereas unlabelled data is cheap and available in abundance. Semi-supervised learning use limited label data to learn the input variable and make some initial predictions required for the unsupervised learning process [5]-[8].

3. Related work

To overcome the problem of unsupervised clustering process, user can provide some priori information regarding the underlying structure of the data. Such information can be made available by the presence of some labelled data along with unlabelled data. Incorporating such information in the clustering process, helps to better guide the process towards better partitioning and avoid local minima. Such process is known as Semi-Supervised clustering process [5].

Most widely used Fuzzy c-means (FCM) given by Bezdek [22] starts with the random initialization of cluster centroids and membership matrix. The process of assignment and updation is completed till some convergence criterion is met. Sometimes FCM algorithm converges to a local solution. Also, the final result of the clustering process highly depends on the initial values of the cluster centroids and membership matrix. To avoid such problem of FCM, the concept of partial supervision was introduced in the process by Pedcryz et al. [23], known as semi-supervised fuzzy c-means.

3.1. Semi-Supervised Fuzzy c-means (SSFCM)

A salient feature of partial supervision in the clustering algorithms is the availability of labelled data [5] or the presence of some constraints [8]. In this SSFCM model partial supervision is provided in the form of labels. Tong et. al. [24] proposed the SSFCM techniques which extends the definition of FCM to capture the hidden and invisible structure of the dataset with the help of partial supervision. The algorithm can be explained as:-

Let data is represented as X with n number of vectors in feature space S^p such that

$X = \{x_1, x_2, x_3, ...x_n\}$

The most basic goal of clustering algorithms is to provide label to the every data point showing its affinity to some class. Let c denotes the number of classes, $1 \le c \le n$. In semi-supervised clustering, a small set of labelled data is provided with the information to encounter some initial variable constraints, along with large set of unlabelled data. The amount of label data available is not sufficient for the process of clustering [8]. So X in case of SSFCM clustering can be represented as

$X = \left\{x_1^l, x_2^l, x_3^l .....x_k^l . \middle| x_1^u, x_2^u, x_3^u ......x_m^u\right\}$ where data with l superscript is labelled and with u superscript is unlabelled.

Consider number of data points in a labelled set as n^l and in unlabelled set as n^u. The total number of data points are denoted as $n = n^l + n^u$. Here membership matrix of labelled data will be set as crisp matrix of 0 and 1, where μ_{ik}^l is set to 1, if x_k is a member of class i and 0 otherwise.

The membership matrix for unlabelled data μ_{im}^u will be initialized randomly [13]. Further initial seed (cluster center) values will be calculated from labelled data and labelled membership matrix as

$$s_i^0 = \frac{\sum_{k=1}^{n^l} \left(\mu_{ik}^l\right)^t x_k^l}{\sum_{k=1}^{n^l} \left(\mu_{ik}^l\right)}, 1 \le i \le c, 1 \le k \le n^l \qquad (2)$$

Further $\left(\mu_{im}^u\right)$ is updated using calculated seed (cluster center) values from Eq. (2) as

$$\left(\mu_{im}^u\right) = \frac{d(x_m^u, s_i)^{1/t-1}}{\sum_{j=1}^{c} d\left(x_m^u, s_j\right)^{1/t-1}}, 1 \le i \le c, 1 \le m \le n^u \qquad (3)$$

where $d(x_m^u, s_i)$ is the Euclidean distance between data point and the seed value. The membership matrix is updated in terms of closeness of an element towards the cluster seed values (Euclidean distance is calculated between every element and cluster centroids). Finally the seed values are updated for complete data set by taking into account the membership of data object with respect to labelled and unlabelled data.

$$s_i = \frac{\sum_{k=1}^{n_l} \left(\mu_{ik}^l\right)^t x_k^l + \sum_{m=1}^{n_u} \left(\mu_{im}^u\right)^t x_m^u}{\sum_{k=1}^{n_l} \left(\mu_{ik}^l\right)^t + \sum_{m=1}^{n_u} \left(\mu_{im}^u\right)^t} \qquad (4)$$

The process of updating the cluster centroids and membership matrix is completed till some defined convergence criterion. In SSFCM, the percentage of unlabelled data is high as compare to labelled data, so some mechanism is required to deal with the problem of uncertainty present in the unlabelled data. It is less robust to noise and outliers.

3.2. Intuitionistic Fuzzy Set (IFS)

Intuitionistic fuzzy set is a higher order fuzzy set and was introduced by Atanassov [25]. Both the membership degree and the non-membership degree are taken into consideration by the intuitionistic set theory. The non-membership degree is taken as the complement of the membership degree in an ordinary fuzzy set while it is taken less than or equal to the complement of the membership degree in an intuitionistic fuzzy set mainly due to the presence of hesitation degree[11].

The Intuitionistic Fuzzy Set can be mathematically represented for dataset X as:

$$R = \left\{x, \mu_R(x), v_R(x) \middle| x \in X\right\} \qquad (5)$$

Here $\mu_R(x) \to [0,1]$, $v_R(x) \to [0,1]$ can be termed as membership and non-membership degree of data point x in an intuitionistic fuzzy set R with the following condition:

$$0 \le \mu_R(x) + v_R(x) \le 1 \qquad (6)$$

Where, if $v_R(x) = 1 - \mu_R(x)$ for data point x in set R, then set R becomes fuzzy set.

The issues associated with the definition of membership and non-membership were resolved with the introduction of IFS theory. It stated that there was still some indecisiveness in defining that up to what measure a data point is associated with a cluster centroid in an intuitionistic fuzzy set R. This indecisiveness (hesitation) arises due to lack of precision in defining the membership degrees which differs on person's choice. It can be generated due to various reasons, for example: inaccuracy in measurement, instrument errors, handling errors, reporting errors and different methods associated with the sampling.

Szmidt and Kacprzyk [26] projected the need of handling this hesitation degree and gave a parameter $\pi_R(x)$(hesitation degree) for better defining the intuitionistic fuzzy index of x for set R:

$$\pi_R(x) = 1 - \mu_R(x) - v_R(x) \tag{7}$$

Here hesitation degree denotes the elements which are neither associated with membership nor non-membership. where $0 \le \mu_R(x) \le 1$

Therefore with $\pi_R(x)$, the Atanassov's intuitionistic fuzzy set is characterized as:

$$R = \{x, \mu_R(x), v_R(x), \pi_R(x) | x \in X\} \tag{8}$$

such as $\mu_R(x) + v_R(x) + \pi_R(x) = 1$

Therefore, to give complete definition of IFS we require to state triple membership functions (i) membership degree $\mu_R(x)$ (ii) non-membership degree $v_R(x)$ and (iii) hesitation degree $\pi_R(x)$. IFS handles better the problem associated with noise and outliers.

4. Proposed Semi-Supervised Intuitionistic Fuzzy c-means (SSIFCM)

The prominent feature of semi-supervised clustering algorithm is that data set X is composed of labeled (x^l) and unlabelled data (x^u). In the proposed SSIFCM, x^l is used to bias the result towards clustering and helps in giving more consistent results. The labeled data will allow the clustering process to specify accurate number of clusters c on the basis of class information available. So the data should be labeled with the given constraint that from every class, some percentage of data should be labeled, in order to provide training patterns that could capture a training set from every class. Table 1 shows the details of the convention symbols used for defining the process.

Table 1. Conventional Symbols used in the Proposed Technique

Symbol	Quantity	Detail
X	Data set	Data set that is to be clustered
x^l	label data	Data used for initiating clustering process.
x^u	Unlabel data	Data to be clustered.
c	Number of clusters	Number in which data has to be classified.
s_i	Seed value	Centre of the clusters.
u_R	Membership matrix	Degree of participation of a point with the cluster.
v_R	Non-membership matrix	Non-degree of Participation
π_R	Hesitation degree	Uncertainty in participation
s_i^*	Updated Centre	Centre including labelled and unlabelled data.
u_R^*	Updated Membership	Intuitionistic membership of unlabelled data.
t	Degree of fuzziness	Generally set to 2.
η, λ	User defined constants	User defined constants for the process of tuning.
ε	Termination constant	Generally set to 0.0001
$d(x_m^u, s_i^*)$	Distance Metrics	Euclidean distance between seed value and data point.

In the proposed SSIFCM, intuitionistic fuzzy set theory is used to handle the problem of uncertainty in the unlabelled data used in semi-supervised clustering process. Intuitionistic fuzzy sets (IFS) are generally defined by triple membership degrees: (i) initial membership degree of unlabelled data is defined using initialized seed values (cluster centroids) from the labelled data as in Eq. (2), where the membership matrix of label data is used to calculate cluster centroids (ii) Intuitionistic fuzzy generators are used to calculate the non-membership degree (Yager's and Sugeno fuzzy generators) (iii) further, hesitation degree is calculated.

Generally in FCM, non-membership is complement of membership, but in IFS, non-membership is tuned using fuzzy generators in order to handle the uncertainty (hesitation degree) present in the data. A general study of fuzzy generators has been done in [27].

In the proposed algorithm, the membership matrix of unlabelled data is calculated using initialized seed values of the clusters from the labelled data as given in Eq. (2).

$$\mu_R\left(x^u\right)=\frac{d(x_m^u,s_i)^{1/t-1}}{\sum_{j=1}^{c} d\left(x_m^u,s_j\right)^{1/t-1}},1\le i\le c,1\le m\le n^u \tag{9}$$

Generally in literature Yagers and Sugeno are used for the purpose of calculating non-membership in Intuitionistic fuzzy sets (IFS).

A non-membership degree can be generated using Sugeno[27] class as

$$v_R\left(x^u\right)=\frac{1-\mu_R\left(x^u\right)}{1+\eta\mu_R\left(x^u\right)},\eta>0 \tag{10}$$

If non-membership degree is generated using Yagers class [28], then it is calculated as

$$v_R\left(x^u\right)=\left(1-\mu_R\left(x^u\right)^\lambda\right)^{1/\lambda},\lambda>0 \tag{11}$$

In the proposed technique we are going to investigate the results using Sugeno negation function. Non-membership matrix is calculated using the Sugeno generator using Eq. (10). Further hesitation degree is calculated as

$$\pi_R(x^u)=1-\mu_R(x^u)-v_R(x^u) \tag{12}$$

An intuitionistic membership matrix is calculated as the sum of membership and hesitation degree for the unlabelled data.

$$\mu_R^*\left(x^u\right)=\pi_R(x^u)+\mu_R\left(x^u\right) \tag{13}$$

Further centers and membership matrix are updated using new intuitionistic membership matrix as

$$s_i^*=\frac{\sum_{k=1}^{n_l}\left(\mu_{ik}^l\right)^t x_k^l+\sum_{m=1}^{n_u}\left(\mu_{im}^{*u}\right)^t x_m^u}{\sum_{k=1}^{n_l}\left(\mu_{ik}^l\right)^t+\sum_{m=1}^{n_u}\left(\mu_{im}^{*u}\right)^t} \tag{14}$$

$$\left(\mu_R^*\right)=\frac{d(x_m^u,s_i^*)^{1/t-1}}{\sum_{j=1}^{c} d\left(x_m^u,s_j^*\right)^{1/t-1}} \tag{15}$$

The process is repeated till the termination criterion ($s_{i+1}^*-s_i^*\le\varepsilon$) is met.

In the proposed Semi- Supervised Intuitionistic Fuzzy c-means algorithm, steps are executed as per the given algorithm SSIFCM.

Proposed Algorithm: SSIFCM

```
Step 1: Define the dataset as labelled and unlabelled.
Step 2: Set the membership of labelled data with 0 and 1 label.
Step 3: Initialize the centres of clusters using label data and membership matrix of label data as in Eq. (2).
Step 4: Repeat steps till the termination criterion is met
        (a) Update the distance matrix for the unlabelled data with corresponding centres using Euclidean distance norm.
        (b) Calculate membership matrix using eq. (9)
        (c) Calculate the non-membership using eq. (10)
        (d) Calculate hesitation present using eq. (12)
        (e) Calculate the intuitionistic membership matrix using eq. (13)
        (f) Update new cluster centres using intuitionistic membership matrix and taking information from labelled and unlabelled data
            using eq. (14).

Step 5: Using the final membership matrix, group the remaining samples of unlabelled data.
```

5. Experimental Results and Discussions

A number of experiments were carried out for a range of data sets using the SSIFCM clustering. The experiments are performed to analyze the results of proposed technique with respect to unsupervised, semi-supervised and supervised learning processes. The main aim of this detailed analysis is to come up with a comparison of the performance of the SSIFCM clustering with the well-known clustering techniques like FCM and SSFCM. In this part, we have conducted the experiments on four real benchmark data sets to check the performance of different clustering algorithms. The benchmark data sets employed are Iris data, Seed data, Abalone data and wine dataset taken from UCI repository [29]. Further experiments are carried on a natural image, embedded with noise to prove the robustness of algorithm in the field of image segmentation. The selection of data sets includes assortments in the number of clusters, number of data points and count of features of each datum. For all data sets, we choose following assumptions as t = 2 which is a common choice for fuzzy clustering, maximum iteration = 100 and termination constant ε = 0.0001. The proposed algorithm is embedded with Sugeno fuzzy generator where the value of η is set for every data on the basis of trial and experimentation. The labelling is done randomly with the assumption that every class should have some labelled data, in order to maintain the consistency of experimentation.

5.1. Benchmark Datasets

Iris data set consist of 3 classes of 50 instances each, where every class classifies a form of iris plant and can be consider as one separate cluster in the experiments. Each instance has 4 features defining petal length, petal width, sepal length and sepal width respectively. In this, one group can be completely separated from the other two groups while group 2 and 3 has some overlaps. Secondly seed data set is comprised of kernel that belongs to 3 different classes of wheat named as Kama, Rosa and Canadian. Each class consists of 70 components selected arbitrarily for the experiment. Each component is defined by seven geometric parameters of wheat kernels. Thirdly Abalone dataset defines the age of abalones, which are large, edible sea snails. Here each instance is defined by 8 attributes. Total numbers of instances are 4177 which are classified into 3 categories. Fourthly Wine dataset comprises of data about the chemical study of 178 wines grown in the same part of Italy but cultivated by three different cultivators. The difficulty is to differentiate different types of breed based on 13 continuous features derived from chemical study. The dataset consist of three clusters with 59, 71 and 48 instances for each cluster. The brief information of these dataset is listed in table 2.

In the experimental study, data is partitioned into different clusters using different clustering and classification algorithms. The results are compared by calculating misclassification error and the clustering accuracy. The results of experiments are assessed using Huang's accuracy measure [17]:

$$r = \frac{\sum_{i=1}^{k} n_i}{n} \tag{16}$$

where n_i represents the true positive of data present in both the i^{th} cluster and simultaneously in its true cluster and n is the total number of data points in the data set. The greater value of accuracy measure r substantiates better clustering results with impeccable clustering giving a value $r = 1.$

Table 2. Brief Information of UCI datasets

Dataset	Data Points	Features	Number of Clusters
Iris	150	4	3
Seed	210	7	3
Abalone	4177	8	3
Wine	178	13	3

In the first experiment, 10% of label data is provided for the process of semi-supervised learning. The labelling of data is done randomly, so for each data set, five trials have been conducted and finally mean of the misclassification error is calculated. Further accuracy based on Huang index as in Eq. (16) is calculated. The results are calculated by our proposed clustering algorithm and compared with an unsupervised learning, supervised learning and semi-supervised learning. The unsupervised learning involves widely used Fuzzy c-means [22], supervised learning involves naïve bayes (NB), k-nearest neighbour and SVM classification algorithms [1] and semi-supervised learning is calculated using SSFCM [13].

Table 3 shows the result of misclassifications obtained in the process of different learning techniques with 10% of label data. It has been observed that supervised learning techniques failed to perform with less label information. In iris and wine data sets, SVM and NB fails to show the results due to unavailability of enough label data, in order to calculate the variance in each class. KNN also shows very poor results, when limited number of label information is available for the process of training.

Table 3. Misclassifications obtained in different leaning techniques in comparison to the proposed SSIFCM technique with 10% label data

Datasets	Count	Label/ Unlabel data	Misclassification					
			FCM	NB	SVM	KNN	SSFCM	SSIFCM (Proposed)
Iris Data Set η =0.83	1	15/135	16	10	8	76	14	8
	2	15/135	16	-	-	55	13	10
	3	15/135	16	-	21	48	12	10
	4	15/135	16	9	12	47	14	11
	5	15/135	16	11	62	49	17	16
	Avg		16	-	-	55	14	**11**
Wine Data Set η =0.86	1	17/161	56	34	62	36	54	52
	2	17/161	56	-	39	18	52	51
	3	17/161	56	16	-	30	46	45
	4	17/161	56	34	49	113	51	49
	5	17/161	56	-	-	15	49	51
	Avg		56	-	-	43	51	**49**
Seed Data Set η =0.81	1	21/189	22	25	24	22	19	17
	2	21/189	22	32	20	39	20	15
	3	21/189	22	30	22	31	19	16
	4	21/189	22	50	25	79	21	20
	5	21/189	22	21	21	22	21	18
	Avg		22	32	23	39	20	**17**
Abalone Data Set η =0.83	1	417/3760	2019	1803	1729	1772	1885	1867
	2	417/3760	2019	1797	1767	1799	1842	1829
	3	417/3760	2019	1835	1770	1759	1839	1828
	4	417/3760	2019	1711	1732	1773	1833	1840
	5	417/3760	2019	1781	1767	1811	1845	1848
	Avg		2019	1785	1753	1783	1849	**1842**

SSFCM has performed well with the availability of limited amount of label data as compared to stated techniques for iris, wine and seed data sets. However for the abalone data, results shown by our proposed algorithm are better as compared to FCM and SSFCM, but results shown by all supervised learning techniques are better than our proposed algorithm.

Table 4 shows the result of the accuracy measure based on the Huang's index on all the four datasets. From table 4, it can be seen that SSFCM performs better than FCM on every dataset but SVM and NB performs well in case of abalone dataset.

The proposed algorithm SSIFCM performs well as compared to FCM, NB, SVM, KNN on the iris, wine and seed dataset. But on abalone dataset, SVM performs relatively well. But if only single label is provided from any class, then supervised learning process fails to train the data for the process of classification and unable to show the results.

From these results, it can be stated that supervised learning techniques perform extremely well, when the complete label data is available in order to train the data properly for the process of classification. In the absence of complete label data, unsupervised and semi-supervised clustering can perform well.

Table 4. Comparison of the proposed technique SSIFCM on the basis of accuracy by Huang's Index

Data sets/ Techniques	Accuracy by Huang's Index					
	FCM	NB	SVM	KNN	SSFCM	SSIFCM (Proposed)
Iris	0.89	-	-	0.63	0.91	**0.93**
Wine	0.69	-	-	0.75	0.71	**0.72**
Seed	0.89	0.84	0.89	0.81	0.90	**0.92**
Abalone	0.51	0.57	0.58	0.57	0.55	**0.56**

Secondly, we performed the detailed analysis of proposed techniques SSIFCM with different clustering techniques. Table 5 shows the comparative analysis of our proposed method with FCM and SSFCM in terms of number of misclassification and accuracy for varying increase in labelled data. The results show that our proposed method shows better results as compared to FCM and SSFCM for all datasets. The semi-supervised approaches perform better with the increase in labelled data. As the amount of label information increases, the degree of hesitation decreases so the SSFCM and proposed SSIFCM tend to perform with the same accuracy. With the increase in the weight of labelled data, clusters centres more accurately point to the more real centroids of the class. Thus the percentage of the labelled data greatly influences the process of clustering. But SSIFCM is found to be more effective and robust as compared to SSFCM with less percentage of labelled data.

To prove the robustness of proposed technique, we have tested the results on different amount of label information. We have also used two additional performance measures to validate our proposed clustering algorithm, known as external validation measures. We have used normalised mutual information (NMI), also known as information theoretic measure and secondly rand index (RI) exist pair-counting-based measure. The closer the value of RI and NMI is towards 1, better is the accuracy of clustering [17].

Table 5. Comparative analysis of SSIFCM algorithm with FCM and SSFCM in terms of number of misclassifications and accuracy.

Labelled Percentage	Unlabelled Data	Misclassification			Accuracy (%)		
		FCM	SSFCM	SSIFCM (Proposed	FCM	SSFCM	SSIFCM (Proposed)
Iris Data Set							
10%	135	17	15	13	0.88	0.90	**0.916**
20%	120	17	14	11	0.88	0.906	**0.92**
30%	105	17	13	10	0.88	0.91	**0.93**
Seed Data Set							
10%	189	22	21	18	0.89	0.90	**0.914**
20%	168	22	17	12	0.89	0.91	**0.94**
30%	147	22	15	11	0.89	0.92	**0.947**
Abalone Data Set							
10%	3760	2019	1852	1841	0.51	0.55	**0.56**
20%	3342	2019	1609	1602	0.51	0.61	**0.62**
30%	2924	2019	1405	1397	0.51	0.66	**0.665**
Wine Data Set							
10%	161	56	54	52	0.69	0.712	**0.72**
20%	143	56	42	38	0.69	0.76	**0.78**
30%	125	56	37	37	0.69	0.79	**0.79**

Table 6. Comparative analysis of SSIFCM algorithm with FCM and SSFCM in terms of Random Index and NMI

Labelled Percentage	Unlabelled Data	Random Index			NMI		
		FCM	SSFCM	SSIFCM (Proposed	FCM	SSFCM	SSIFCM (Proposed)
Iris Data Set							
10%	135	0.67	0.69	**0.71**	0.69	0.71	**0.73**
20%	120	0.67	0.71	**0.73**	0.69	0.72	**0.74**
30%	105	0.67	0.77	**0.77**	0.69	0.76	**0.76**
Seed Data Set							
10%	189	0.70	0.71	**0.72**	0.66	0.67	**0.69**
20%	168	0.70	0.72	**0.76**	0.66	0.68	**0.72**
30%	147	0.70	0.73	**0.77**	0.66	0.70	**0.732**
Abalone Data Set							
10%	3760	0.14	0.143	**0.148**	0.15	0.157	**0.158**
20%	3342	0.14	0.155	**0.158**	0.15	0.162	**0.163**
30%	2924	0.14	0.171	**0.172**	0.15	0.167	**0.168**
Wine Data Set							
10%	161	0.34	0.35	**0.36**	0.69	0.712	**0.72**
20%	143	0.34	0.352	**0.367**	0.69	0.76	**0.78**
30%	125	0.34	0.37	**0.37**	0.69	0.79	**0.79**

Table 6 illustrates the results on external validation indices (RI and NMI) for different clustering algorithms in comparison to the proposed SSIFCM. It is observed that the performance of the proposed semi- supervised clustering algorithm is better or comparable with other clustering algorithms.

5.2. Results on Images

In this section, we test our algorithm on a natural image to prove the efficacy of the proposed algorithm with respect to the robustness to the noise and segmentation accuracy. Image is basically a fuzzy dataset where the pixels are imprecisely defined. It contains lot of uncertainty and inhomogeneity. The proposed technique is shown to work well on noisy images. To prove the efficacy of SSIFCM, the results are compared with FCM and SSFCM. The performance is evaluated on the basis of segmentation accuracy where the key goal is to measure the correctness of clustering.

The segmentation accuracy is basically defined as assigning pixels coming from the testing pool to the correct cluster, that is, how often the clustering process meets the actual assignment given a ground truth values[17]. The performance parameter SA is:

$$SA = (\text{Number of correctlyclassifiedpixels})/(\text{Total Number of pixels}) \tag{17}$$

An assignment is accurate if the pixel from the sample image is assigned to one of the clusters of the correct class (i.e. the target). The closer the value of SA towards 1, better the accuracy achieved. A natural image of moon is taken with 350X442X3 size, and clustered into 2 segments. Image is further added with 10% of Gaussian noise, in order to prove robustness of proposed technique for the segmentation of noisy image.

Table 7. Comparative analysis of SSIFCM algorithm with FCM and SSFCM in terms of Segmentation Accuracy

Noise	Labelling	Segmentation Accuracy		
		FCM	SSFCM	SSIFCM (Proposed)
10%	10%	0.7313	0.7501	**0.7789**
	20%	0.7313	0.7712	**0.7815**
	30%	0.7313	0.8016	**0.8417**
20%	10%	0.7215	0.7412	**0.7605**
	20%	0.7215	0.7625	**0.7791**
	30%	0.7215	0.8016	**0.8414**

Fig. 3 shows the result of experiment performed on natural image. Fig. 3(a) shows the result of FCM where the first image shows original image, second image shows the result of overall segmentation of image with two clusters followed by different object in the image.

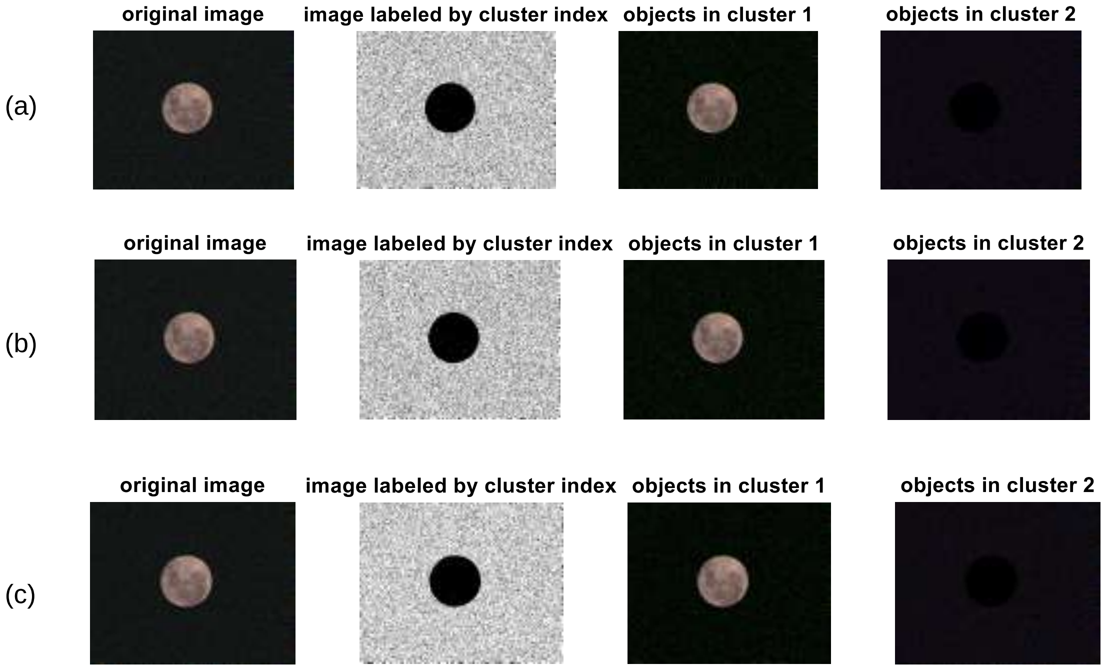

Figure 3. Natural image of moon with 2 clusters (a) result of FCM (b) result of SSFCM with 10% labelling (c) result of SSIFCM with 10% labelling.

Fig 3(b) shows the results of SSFCM and Fig 3(c) shows the result of proposed technique and shows to give better results.

Table 7 gives the comparative results of segmentation accuracy of FCM, SSFCM with the proposed technique SSIFCM. Overall segmentation accuracy is calculated for the process of clustering on the image corrupted with different percentage of Gaussian noise. The results are calculated for different percentage of labelled data as the input to semi-supervised clustering process. From the table, it is observed that segmentation accuracy increases with the increase in the labelled pixels. The segmentation accuracy increases from 77% to 84% with the increase in the labelling of pixels.

5.3. Limitations

The propose technique has certain limitation as the result varies with change in the label information. Every class should be provided with at least one label, in order to provide some representative structure to each class. The parameter η of intuitionistic fuzzy set need to be tuned manually for each data set.

6. Conclusion

The proposed algorithm SSIFCM is an intuitionistic approach towards the process of semi-supervised clustering technique. The proposed algorithm is compared with FCM, SSFCM and some supervised learning algorithms on benchmark data sets. The proposed technique SSIFCM proved to give good result even with small amount of supervision only. It has been shown that when 10% or 20% of the data is labelled only, the results shown by the proposed algorithm are superior to all other stated algorithms. This is due to the inclusion of intuitionistic approach in the semi-supervised approach. As the labeling percentage increases to 30%, the uncertainty decreases and the proposed algorithm shows comparable results to semi-supervised or supervised approaches. The proposed SSIFCM proved to give robust results for image segmentation for noisy images with the accuracy of 77.89% in comparison to SSFCM with an accuracy of 75.01%. This shows that our proposed algorithm can obtain better performance with limited amount of supervision available.

References

[1] Sharma, M., Singh, G. and Singh, R. (2017) Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. *Elsevier Manon IRBM* 38 : 305-324.

[2] Williams, N., Zander, S. and Armitage, G. (2006) A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. *ACM SIGCOMM Computer Communication Review* 36(5) : 7-15.

[3] Ren, T., Wang, H., Feng, H., Xu, C., Liu, G. and Oing, P. (2019) Study on the improved fuzzy clusteing algorithm and its application in brain image segmentation. *Applied Soft Computing Journal* 81: 105503.

[4] Chen, X., Yu,G., Tan, Q. and Wang, J. (2019) Weighted Samples based semi-supervised classification. *Applied Soft Computing Journal* 79: 46-58.

[5] Basu, S., Banerjee, A. and Mooney, R. (2002) Semi-supervised clustering by seeding. *Proceedings of the Int. Conference on Machine Learning* : 19–26.

[6] Demiriz, A., Bennett., K. and Embrechts, M. (1999) Semi-supervised clustering using genetic algorithms. *Intelligent Engineering Systems*: 809–814.

[7] Blum, A. and Chawla. S. (2001) Learning from labelled and unlabelled data using graph mincuts, *Proceedings in 18th International Conference in Machine Learning*.

[8] Dinler, D. and Tural, M.K. (2017) Robust semi-supervised clustering with polyhedral and circular uncertainty. *Neurocomputing* 265: 4-27.

[9] Saha, S., Alok, A.K. and Ekbal, A. (2016) Brain image segmentation using semi-supervised clustering. *Expert system with Application* 52: 50-63.

[10] Sharma, M., Sharma, S. and Singh, G. (2018) Performance Analysis of Statitical and Supervised Learning techniques in Stock Data Mining. *MDPI Journal* 3(54): 1-16.

[11] Bensaid, A.M., Hall, L.O., Bezdek, J.C. and Clarke, L.P. (1996) Partial Supervised Clustering for Image Segmentation. *Pattern Recognition* 29 : 859-871.

[12] Pedrycz, W. and Waletzky, J. (1997) Fuzzy Clustering with partial supervision. *IEEE Trans. on Systems, Man, and Cybernetics, Part B-Cybernetics* 27 : 787-795

[13] Zhang, D., Keren, T., and Chen, S. (2004) Semi-supervised Kernel Based Fuzzy C-Means. *ICONIP, LNCS.* 3316: 1229-1234

[14] Bennett, K. and Demiriz, A. (1999) Semi-Supervised Support Vector Machines. *Advances in Neural Information Processing Systems* 11 : 368-374.

[15] Dubey, Y. K., Mushrif, M.M. and Mitra, K.(2016) Segmentation of brain MR images using rough set based intuitionistic fuzzy clustering. *Biocybernetics and Biomedical Engineering* 36: 413-426.

[16] Arora, J. and Tushir, M. (2018) Robust spatial intuitionistic fuzzy C-means with city-block distance clustering for image segmentation. *Journal of Intelligent & Fuzzy Systems* 35: 5255-5264.

[17] Arora, J. and Tushir, M. (2017) A new kernel-based possibilistic intuitionistic fuzzy c-means clustering. *International Journal of Artificial Intelligence and Soft Computing* 6(4) : 306-325.

[18] Xu, Z. and Wu, J. (2010) Intuitionistic Fuzzy Clustering Algorithms. *Journal of Systems Engineering and Electronics* 21(4) : 580-590

[19] Pelekis, N., Iakovidis, D. K., Kotsifakos, E. E. and Kopanakis, I. (2008) Fuzzy clustering of intuitionistic fuzzy data. *International Journal of Business Intelligence and Data Mining* 3(1) : 45-65.

[20] Chaira, T. (2010) A novel Intuitionistic fuzzy c means clustering algorithm and its application to medical images. *Applied Soft Computing* 11(2) : 1711-1717.

[21] Smith, M. R. and Martinez, T. (2011) Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified. *Proceedings of International Joint Conference on Neural Networks (IJCNN)* 2690–2697.

[22] Ruspini, E. H., Bezdek, J. C. and Keller, J. M. (2019) Fuzzy Clustering: A Historical Perspective. *IEEE Computational Intelligence Magazine* 14(1): 45-55.

[23] Pedrycz, W. and Waletzky, J. (1997) Fuzzy Clustering with partial supervision. *IEEE Trans. on Systems, Man, and Cybernetics, Part B-Cybernetics* 27 : 787-795.

[24] Tong, X., Jiang, Q., Sang, N., Gan, H. and Zeng, S. (2009) The Feature weighted FCM algorithm with semi-supervised. *Proceedings of Eighth International Symposiumon Distributed Computingand Applications to Business, Engineering and Science* : 22-26.

[25] Atanassov, K. (1986) Intuitionistic fuzzy sets. *Fuzzy Sets Systems* 20(1) : 87–96

[26] Szmidt, E. and Kacprzyk, J. (2000) Distances between intuitionistic fuzzy sets. *Fuzzy sets Systems* 114.

[27] Bustince, H. and Mohedano, V. (1997) About the Intuitionistic Fuzzy Set Generators. *First Int. Conf. on IFS, Sofia, NIFS* 3(4): 21-27.

[28] Yager, R. R. (2009) Some aspects of Intuitionistic fuzzy sets. *Fuzzy Optimization and Decision making* 8(1): 67-90.

[29] UCI Repository of Machine Learning Databases, University of California, Irvine, available from: http://www.ics.uci.edu/~mlearn/MLRository.html