ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
www.ijptonline.com
Research Article

CLUSTERING CATEGORICAL DATA USING INTUITIONISTIC FUZZY K-MODE ALGORITHM

B.K. Tripathy[1], Akarsh Goyal*[2], Patra AnupamSourav[3]
School of Computer Science and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.
Email: akarsh.goyal15@gmail.com

Received on 10-08-2016 *Accepted on 30-08-2016*

Abstract

In this paper we introduce the concept of intuitionistic fuzzy k-mode algorithm to cluster categorical data. This notion is an extension of fuzzy k-mode in which we have added a new parameter known as intuitionistic degree in the calculation of membership values of element x in a given cluster. Systematic experiments were carried out with datasets taken from the UCI Machine learning repository. The results and a comparative evaluation show a high performance and consistency of the proposed method, which achieves significant improvement compared to fuzzy k-mode. Intuitionistic fuzzy k-mode is very efficient when clustering large data sets, which is very much critical to data mining applications.

Keywords: Categorical data, Clustering, Data mining, Fuzzy k-mode, Intuitionistic fuzzy k-mode.

1. Introduction

Data Mining is the computational process of finding patterns in large data sets involving methods at the which belong at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and convert it into an understandable structure for further use. Many techniques are used to extract valid patterns and knowledge mining from complex and huge amount of data set, such as association, classification, clustering, pattern recognition etc. which are used to group, or classify the dataset. In this paper we will be working on clustering algorithm for mining purposes. Clustering is the job of grouping a set of objects in a way such that objects in the same group are more similar to each other than to those in other groups [7]. Hence a cluster is a group of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Nowadays most of the raw data available is without any class values in which the different records can be classified or without much relation to each other. So in these cases the concept of clustering comes in handy. Clustering methods are used to minimize the inter cluster similarity and maximize the intra cluster similarity. Categorical data is the statistical data type consisting of variables that can take on one of a limited, and usually fixed, number of possible values, thus assigning each individual to a particular group or "category." The objects in the data base contain the attributes of various data types. These values may be of either numeric or non-numeric type. Categorical dataset therefore generally involves nominal, ordinal and interval-scaled attributes as shown in figure below.

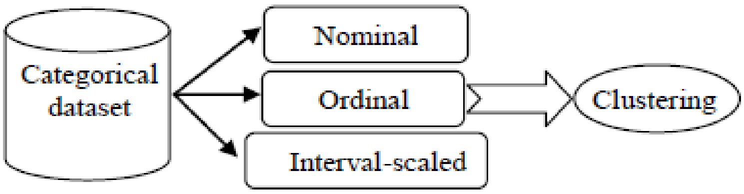

Fig 1: Clustering Categorical Data.

Clustering can be performed for both numerical and categorical data and for clustering numerical data. But clustering categorical data is very different and difficult from those of numerical data. The distance metric can't be applied to the categorical data directly. So k-means algorithm which is the most used clustering method is rendered ineffective when applied on categorical data. This is because it wholly depends on the distance metric and it can only minimize a numerical cost function. So for categorical data we have to use k-mode method. The k-modes[5]approach modifies the standard k-means process for clustering categorical data by substituting the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centers and updating modes with the most frequent categorical values in each of iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minima result. Fuzzy k-modes[6] is an extension of k-modes. In this method a fuzzy partition matrix is generated from categorical data within the framework of the fuzzy k-means algorithm[8]. Its main objective is to provide a method to find the fuzzy cluster modes when the simple matching dissimilarity measure is used for categorical objects. The fuzzy version has improved the -modes algorithm by assigning confidence to objectsin different clusters. In this paper we have devised a new concept called intuitionistic fuzzy k-modes. It is an addendum of fuzzy k-mode. In 1986 Atanassov developed the intuitionistic fuzzy set[1] concept. For all IFSs, Atanassov also indicated an intuitionistic degree, $\pi_A(x)$, which arises due to lack of knowledge in defining membership degree. So here we will be using the notion of intuitionistic fuzzy k-modes to do clustering[4]on categorical data to get better outcome than obtained on applying fuzzy k-mode.

2. Related Work

There has been immense amount of work that has been done in the field of data mining and data clustering. Since there is no fixed method of clustering data, new methods have been proposed frequently. Let us first know the history of clustering. In [7] an iterative technique of partitioning a dataset into C-clusters was introduced by McQueen in 1967. Similarly the fuzzy set theory was introduced by Lotfi A. Zadeh in [9] .Applying this concept on clustering Ruspini first proposed the fuzzy clustering algorithm mentioned in [8], which was later modified and generalized by Dunn and Bezdek respectively in [2]. The concepts of k-mode and fuzzy k-mode were introduced by Z. Huang in [5] and [6] respectively. In 1986 Atanassov K T. developed the intuitionistic fuzzy set theory written in [1], on the basis of which Chaira T. formulated the intuitionistic fuzzy clustering algorithm [3], [4].The details of all the algorithms have been discussed in the forthcoming sections of the document.

3. Datasets Used

The datasets used in this paper was taken from UCI dataset repository where various datasets are available for public use. The datasets used are soybean dataset, iris dataset and wine dataset. The description for these datasets is given in the table below.

- Table 1 gives the description of the different datasets used by us.

Table 1: Datasets Description.

Data Set ⟶	Soybean Dataset	Wine Dataset	Iris Dataset
Characteristics	Multivariate	Multivariate	Multivariate
Attribute Type	Categorical	Real, Integer, Categorical	Real, Categorical
Associated Tasks	Classification	Classification	Classification
Number of Instances	47	178	150
Number of Attributes	35	13	4
Missing Values	No	No	No
Class Values	D1,D2,D3,D4	1-3	Iris Setosa, Iris Versicolour, Iris Virginia

4. Notation

In this section we have explained the notations which have been used to give the various equations. The notations relating to categorical data and intuitionistic fuzzy k-mode have been provided.

4.1 Categorical Data

We assume that a database T-stores the set of objects to be clustered defined by a set of attributes $A_1,\ A_2 \ldots\ A_m$. Each attribute A_j describes a domain of values denoted by DOM(A_j) and is associated with a defined semantic and a data type. In this letter, we only consider two general data types, numeric and categorical and assume other types used in database can be linked with one of these two types. The domains of attributes associated with these two types are called numeric and categorical, respectively. A numeric domain consists of real numbers. A domain DOM(A_j) is defined as categorical if it is finite and unordered, e.g., for any a,b $\in$DOM(A_j) either a=b or $a \neq b$.A conjunction of attribute-value pairs logically represents an object X in T as follows: $[A_1 = x_1] \wedge [A_2 = x_2] \wedge \ldots \wedge [A_m = x_m]$, where $x_j \in$DOM(A_j) for $1 \leq j \leq m$.

Without ambiguity, we represent X as a vector$[x_1, x_2, x_3, ..., x_m]$. X is called a categorical object if it has only categorical values. We consider every object has exactly m attribute values. If the value of an attribute A_j is missing, then we denote the attribute value of A_j by null. Let X = {X_1, X_2,… X_n} be a set of n objects. Object X_i is represented as [x_{i1}, x_{i2},. . . , x_{im}]. We write $X_i = X_k$ if $x_{i,j} = x_{k,j}$ for 1 $\leq$j $\leq$m. The relation $X_i = X_k$ does not mean that X_i and X_k is the same object in the real world database. It means the two objects have equal values for the attributes A_1; A_2;. . . . ; A_m.

4.2 Intuitionistic Fuzzy Set

The notion of intuitionistic fuzzy sets introduced by Atanassov[1] emerges from simultaneous consideration of membership values m and non-membership values n of elements of a set. An IFS A in X is given as $\{(x, m_A(x), n_A(x)) |\ x \in X\}$, where $m_A : X \to [0,1]$ and $n_A : X \to [0,1]$ such that $0 \leq m_A(x) + n_A(x) \leq 1$ where $\forall x \in X$. $m_A(x)$ and $n_A(x)$ are membership and non-membership values of an element x to set A in X. Set A becomes a fuzzy set when $n_A(x) = 1 - m_A(x)$ for every x in set A. For all IFSs, Atanassov also indicated an intuitionistic degree, $\pi_A(x)$. This arises due to lack of knowledge in defining membership degree, for each element x in A and this is given as

$$\pi_A(x) = 1 - m_A(x) - n_A(x),\ 0 \leq \pi_A(x) \leq 1 \qquad (1)$$

Membership values $m_A(x)$ lie in an interval range $[m_A(x) - \pi_A(x),\ m_A(x) + \pi_A(x)]$ due to hesitation degree,

Construction of Intuitionistic Fuzzy Set (IFS) is done from intuitionistic fuzzy generator (IFG). In this study, Sugeno's IFG is used. Sugeno's intuitionistic fuzzy complement is written as

$$N(m(x)) = (1 - m(x)) \ / \ (1 + \lambda\, m(x))\ \lambda > 0,\ N(1) = 0,\ N(0) = 1 \qquad (2)$$

Sugeno type intuitionistic fuzzy complement N(m(x)) is used to calculate non-membership values. With Sugeno type fuzzy complement, the hesitation degree is given by

$$\pi_A(x) = 1 - m_A(x) - (1 - m_A(x)) \ / \ (1 + \lambda\, m_A(x)). \qquad (3)$$

5. Methods and Algorithms

5.1 Distance Function

Between two objects *X* and *Y* described by *m* categorical attributes the distance function in k-modes is calculated as

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (4)$$

where

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j; \\ 1, & x_j \neq y_j. \end{cases}$$

Here, x_j and y_j are the values of attribute j in X and Y. This equation is often referred to as simple matching dissimilarity measure or Hemming distance. The larger the number of mismatches of categorical values between X and Y is, the more dissimilar the two objects.

5.2 K-modes algorithm (KM)

In general, k-modes clustering[5] can be expressed as an optimization process of partitioning a data set *D* into *k* clusters by iteratively finding *W* and *Z* that minimize the cost function:

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \omega_{li} \ d(Z_{l,} X_i) \, (5)$$

Subject to $\omega_{li} \in \{0,1\}, 1 \le l \le k, 1 \le i \le n$ (6)

$$\sum_{l=1}^{k} \omega_{li} = 1, \ 1 \le i \le n \qquad (7)$$

and $0 < \sum_{i=1}^{n} \omega_{li} < n, 1 \le l \le k$ (8)

where $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,m}\}$ denotes m categorical attributes and $D = \{X_1, X_2, \cdots, X_n\}$ represents a set of n categorical objects .The current cluster membership of an object is represented by $W = [w_{li}]$ which is a {0,1} matrix . $Z = [Z_1, Z_2, \cdots, Z_k]$ represents the cluster modes where k is the number of target clusters. k is predetermined before the clustering process starts. The dissimilarity function is used here which has been defined in(2).

To cluster a categorical data set ***X*** into *k* clusters, the k-modes clustering process consists of the following steps:

Step 1: k unique objects are randomly selected as the initial cluster centers (modes).

Step 2: The distances between each object and the cluster mode is calculated and the object is assigned to the cluster whose centre has the shortest distance to the object. This step is repeated until all objects are assigned to clusters.

Step 3: A new mode for each cluster is selected and compared with the previous mode. If it is different, then we go back to Step 2; otherwise, we stop.

K-modes objective function is minimized by this clustering process:

$$F(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{i,l}\,d(x_{ij},z_{lj})$$

where U= [u_{ij}] is an $n\times k$ partition matrix, $Z=\{Z_1, Z_2, \ldots, Z_k\}$ is a set of mode vectors and the distance function $d(.,.)$.

5.3 Fuzzy k-modes algorithm (FKM)

The fuzzy k-modes algorithm[6]was proposed by Huang and Ng for clustering categorical objects. It is based on the extensions to the fuzzy k-means algorithm. The k-modes algorithm is improved by this method as it assigns membership degrees to datain different clusters.In the fuzzy k-modes algorithm, data D is grouped into k clusters by minimizing the cost function (3-6)

Subject to
$$\begin{cases} 1, & \text{if } X_i = Z_l; \\ 0, & \text{if } X_i = Z_h, h \neq l; \\ 1/\sum_{j=1}^{k}\left[\dfrac{d(Z_l,X_i)}{d(Z_h,X_i)}\right]^{\frac{1}{(\alpha-1)}}, & \text{if } X_i \neq Z_l \text{and } X \neq Z,\ 1\le h\le k. \end{cases} \quad (9)$$

where α is the weighting component, $W=(\omega_{li})$ is the $k\times n$ fuzzy membership matrix.

The cluster centers at each iteration are updated by the fuzzy k-mode algorithm by measuring the distance between each cluster centroid and each object. Here, let X and Y be two categorical objects represented by $[x_1, x_2, ..., x_m]$ and $[y_1, y_2, ..., y_m]$, respectively.

The equation is same as (2).

5.4 Intuitionistic Fuzzy K-mode Algorithm (IFKM)

The Intuitionistic fuzzy k-modes proposed by usfollow from intuitionistic fuzzy set.It brings in to account a new parameter that helps in increasing the accuracy of clustering. This parameter is known as the hesitation value and denoted by π.

1. Assign initial cluster centers or modes for c clusters.
2. Calculate the distance *d*between data objects x_k and centroids v_i.
3. Generate the fuzzy partition matrix or membership matrix U as shown by (7).
4. Compute the hesitation matrix π using

   $\pi_A(x) = 1 - \mu_A(x) - \frac{1-\mu_A(x)}{1+\lambda\mu_A(x)} \mid x \in X(10)$

5. Compute the modified membership matrix U' using

   $\mu_{ik}^{'} = \mu_{ik} + \pi_{ik}(11)$

6. The x_k with higher relative frequency of categorical attributes is chosen to be the new representative, i.e. centre or mode
7. Calculate new partition matrix by using step 2 to 5
8. If $\left\|U'^{(r)} - U'^{(r+1)}\right\| < \varepsilon$ then stop else repeat from step 4.

6. Criteria to be used for evaluation

The Davis-Bouldin (DB) and Dunn (D) indexes are one of the most basic performance analysis indexes. They help in evaluating the efficiency of clustering. The results are dependent on the number of clusters one requires.

6.1 Davis-Bouldin (DB) Index

The DB index is defined as the ratio of sum of within-cluster distance to between-cluster distance. It is formulated as given.

$$\mathrm{DB} = \frac{1}{\mathrm{c}}\sum_{\mathrm{i}=1}^{\mathrm{c}} \max_{\mathrm{k}\neq\mathrm{i}}\left\{\frac{\mathrm{S(v_i)+S(v_k)}}{\mathrm{d(v_i,v_k)}}\right\} \quad \text{for } 1 < i,k < c \quad (12)$$

The aim of this index is to minimize the within cluster distance and maximize the between cluster separation. Therefore a good clustering procedure should give value of DB index as low as possible.

6.2 Dunn (D) Index

The D index is similar to DB index. It is used for the identification of clusters that are compact and separated. It is computed by using

$$\mathrm{Dunn} = \min_{\mathrm{i}}\left\{\min_{\mathrm{k}\neq\mathrm{i}}\left\{\frac{\mathrm{d(v_i,v_k)}}{\max_{\mathrm{l}}\mathrm{S(v_l)}}\right\}\right\} \quad \text{for } 1 < k,i,l < c \quad (13)$$

Maximizing the between-cluster distance and minimizing the within-cluster distance is its aim. Hence a greater value for the D index proves to be more efficient.

6.3 Clustering Accuracy

A clustering result can be measured by the clustering accuracy defined as:

$$r = \frac{\sum_{l=1}^{k} a_l}{n} \quad (14)$$

where a_i is the number of instances occurring in both cluster l and its corresponding class and n was the number of instances in the data set. In our numerical tests k is the number of clusters.Hence a greater value of the accuracy means the given method is much better.

7. Results and Discussion

To evaluate the performance and efficiency of the intuitionistic fuzzy k-modes algorithm and compare it with the fuzzy k-modesalgorithm we carried out several tests of these algorithms.

The datasets used were the soybean dataset, iris dataset and wine dataset.We have taken all the three datasets directly from UCI repository. We have not made any changes to the datasets like removing some redundant rows, cleaning the data or removing some attributes. We chose these datasets to test these algorithms because all attributes of the datasets can be treated as categorical.

For the dataset we used the two clustering algorithms to cluster it. For the intuitionistic fuzzy k-modes algorithm we specified λ= 2. The record was assigned X_i was assigned to the

Lth cluster if $\omega_{li} = \max_{1 \le h \le k}\{\omega_{hi}\}$.

If the maximum was not unique, then Xi was assigned to the cluster of first achieving the maximum. We have taken 4,3 and 3 as the number of clusters for soybean, iris and wine dataset respectively. The table below gives the modes of these clusters produced by the two algorithms. The modes obtained with the two algorithms are not identical. This indicates that the intuitionistic fuzzy k-modes and fuzzy k-modes algorithms indeed produce different clusters.

7.1 Modes of the Clusters

In this section we compute the cluster centres for fuzzy k-mode and intuitionistic fuzzy k-modes for three data sets; soybean dataset, iris dataset and wine dataset to show the superiority of intuitionistic fuzzy k-mode algorithm over the fuzzy k-mode algorithm.

7.1.1 Soybean dataset

Tables 2 and 3 show the results obtained by using the Fuzzy k-mode algorithm.

- Table 2shows the cluster centres from column 1 to 22 on applying Fuzzy k-mode to Soybean dataset.
- Table 3shows the cluster centres from column 23 to 36 on applying Fuzzy k-mode to Soybean dataset

Tables 4 and 5 show the results for the Intuitionistic fuzzy k-mode algorithm.

- Table 4shows the cluster centres from column 1 to 22 on applying Intuitionistic Fuzzy k-mode to Soybean dataset.
- Table 5shows the cluster centres from column 23 to 36 on applying Intuitionistic Fuzzy k-mode to Soybean dataset.

Table 2: Columns 1 through 22 for Fuzzy K-Mode.

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0	1	2	1	0	3	1	1	0	2	1	1	0	2	2	0	0	0	1	0	1	1
2	0	1	2	1	0	3	1	1	0	2	1	1	0	2	2	0	0	0	1	0	1	2
3	3	0	2	1	0	3	1	1	1	2	1	1	0	2	2	0	0	0	1	0	1	2
4	5	0	2	1	0	3	1	1	1	2	1	1	0	2	2	0	0	0	1	1	3	0

Table 3: Columns 23 through 36 for Fuzzy K-Mode.

Z_i	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	0	1	0	0	0	3	4	0	0	0	0	0	0	4
2	0	1	0	0	0	3	4	0	0	0	0	0	0	4
3	0	1	0	0	0	3	4	0	0	0	0	0	0	4
4	1	1	0	0	0	0	4	0	0	0	0	0	0	1

Table 4: **Columns 1 through 22for Intuitionistic Fuzzy K-Mode.**

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	3	0	2	1	0	2	0	1	0	1	1	1	0	2	2	0	0	0	1	1	3	0
2	2	0	2	0	0	1	2	1	0	2	1	0	0	2	2	0	0	0	1	0	1	0
3	3	1	1	1	0	1	2	2	0	0	1	0	0	2	2	0	0	0	1	1	3	0
4	5	0	2	1	0	3	1	1	1	2	1	1	0	2	2	0	0	0	1	1	3	0

Table 5: Columns 23 through 36for Intuitionistic Fuzzy K-Mode.

Z_i	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	1	1	0	0	0	0	4	0	0	0	0	0	0	1
2	1	0	1	2	1	0	4	0	0	0	0	0	0	2
3	1	1	1	0	0	3	4	0	0	0	0	0	1	1
4	1	1	0	0	0	0	4	0	0	0	0	0	0	1

7.1.2 Iris Dataset

Table 6 shows the results obtained by using the Fuzzy k-mode algorithm and Intuitionistic Fuzzy k-mode algorithm.

- Table 6 shows the cluster centres on applying Fuzzy k-mode and Intuitionistic Fuzzy k-mode to Iris Dataset.

Table 6: **Columns 1 through 4.**

	Fuzzy K-mode				Intuitionistic Fuzzy K-mode			
Z_i	1	2	3	4	1	2	3	4
1	5	3	1.4	0.2	4.3	2	1.1	0.6
2	5	3	1.5	0.2	7	4.1	1	0.6
3	5	3	1.5	0.2	7	2	3.6	0.5

7.1.3 Wine Dataset

Tables 7 and 8 show the results obtained by using the Fuzzy k-mode algorithm and Intuitionistic Fuzzy k-mode algorithm respectively.

- Table 7shows the cluster centres on applying Fuzzy k-mode to Wine Dataset.
- Table 8shows the cluster centres on applying Intuitionistic Fuzzy k-mode to Wine Dataset.

Table 7: Columns 1 through 13 for Fuzzy K-mode algorithm.

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13
1	12.37	1.73	2.3	20	88	2.2	2.65	0.43	1.35	3.8	1.04	2.87	520
2	13.05	1.73	2.3	20	88	2.2	2.65	0.26	1.35	2.6	1.04	2.87	680
3	13.05	1.73	2.28	20	88	2.2	2.65	0.43	1.35	4.6	1.04	2.87	680

Table 8: Columns 1 through 13 for Intuitionistic Fuzzy K-mode algorithm.

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.012	0.001	0.0027	0.024	0.07	0.0034	0.003	0.0001	0.0024	0.017	0.001	0.031	1.51
2	0.0142	0.002	0.0025	0.017	0.13	0.0033	0.003	0.0003	0.0004	0.0067	0.0008	0.002	1.45
3	0.0133	0.001	0.0019	0.011	0.07	0.0016	0.001	0.0003	0.0019	0.0029	0.0007	0.004	0.672

7.2 DB and D-index Values

Now we have calculated the DB and D-index of the two algorithms. The representation for this has been made with the help of a table shown below and bar-graphs which clearly indicate that intuitionistic fuzzy k-mode is better than fuzzy k-mode.

- Table 9shows the DB and D index values.
- Figure 2 shows graph of DB for Soybean   Figure 3 shows graph of DB for Iris
- Figure 4 shows graph of DB for Wine   Figure 5 shows graph of D for Soybean
- Figure 6 shows graph of D for Iris   Figure 7 shows graph of D for Wine

Table 9: DB and D-index values.

	Intuitionistic Fuzzy K-mode		Fuzzy K-mode	
Datasets	DB	D	DB	D
Soybean	6.4607	0.2105	12.2026	0.0556
Iris	2.667	0.75	6.5634	0.25
Wine	2.1667	0.9231	7.112	0.2308

Fig 2: Graph of DB for Soybean Dataset.

Fig 3: Graph of DB for Iris Dataset.

Fig 4: Graph of DB for Wine Dataset.

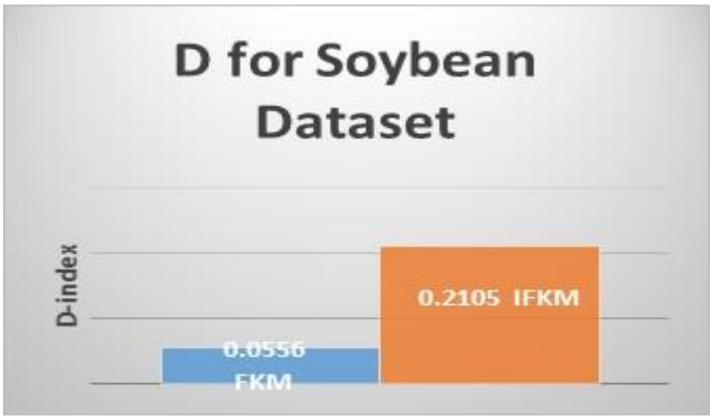

Fig.5: Graph of D for Soybean Dataset.

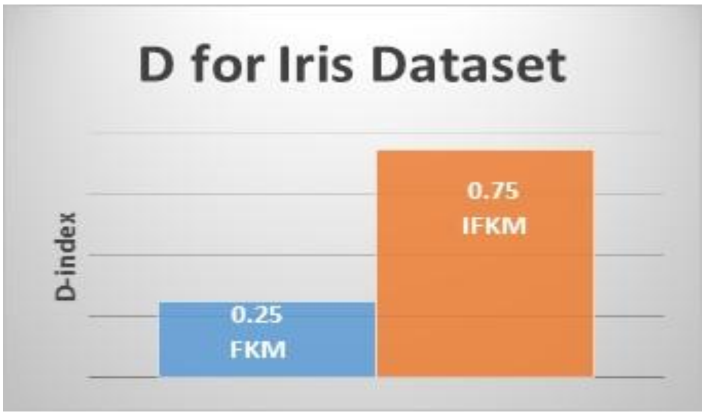

Fig.6: Graph of D for Iris Dataset.

Fig.7: Graph of D for Wine Dataset.

7.3 Accuracy

Now we have calculated the accuracy of clustering of the two algorithms. The accuracies are as follows:

- Table 10 gives the accuracy of clustering.

According to Table 10 the final outcome is true.The accuracy results obtained clearly justify that intuitionistic fuzzy k-mode is a much better algorithm for clustering categorical data than fuzzy k-mode.

Table 10: Accuracy of Clustering.

Datasets	Fuzzy K-mode	Intuitionistic Fuzzy K-mode
Soybean	0.314	0.608
Iris	0.388	0.555
Wine	0.538	0.624

8. Conclusions

Categorical data have become necessary in the real-world databases. However, few efficient algorithms are available for clustering massive categorical data. The development of the k-modes type algorithm and the introduction of fuzzy -modes algorithm for clustering categorical objects based on extensions to the fuzzy –means algorithm was motivated to solve this problem. So we proposed the intuitionistic fuzzy k-mode method in which the intuitionistic degree was taken into effect. This degree leads to an uncertainty in the membership of an object in a particular cluster by a particular value. The complexity of the method remains linear with the additional computation required in the iterative elimination process. The experiments with three commonly referenced data sets from UCI Machine learning repository have shown that the method performs well by using a larger number of initial modes without a need of optimal mode initialization relying on prior knowledge of the data. From the obtained results we perceive that the intuitionistic fuzzy k-mode algorithm performs better than the fuzzy k-mode algorithmas demonstrated in this paper. Information obtained from it is extremely useful in applications such as data mining in which the uncertain boundary objects are sometimes more interesting than objects which can be clustered with certainty.

9. Scope for future Work

We can form better clusters by using a much better distance function. The cluster formed depends heavily on initial cluster we take. Thus finding a way to choose better initial cluster can lead to better cluster formation. Also difference threshold value provides different set of cluster. So according to our application it can be changed for better result.

References

1. Atanassov, K., “Intuitionistic Fuzzy Sets”, *Fuzzy Sets and Systems, Vol. 20*,(1986), pp.87-96.
2. Bezdek, J.C., “Pattern Recognition with Fuzzy Objective Function Algorithms”, Kluwer Academic Publishers, (1981).
3. Chaira, T., “A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images”,*Applied Soft Computing, Vol. 11*, Issue 2, (2011), pp.1711-1717.
4. Chaira, T. and Panwar, A., “An Atanassov’s intuitionistic fuzzy kernel clustering for medical image segmentation”, *International Journal of Computational Intelligence Systems, Vol. 7*, Issue 2,(2014), pp.360-370.
5. Huang, Z., “Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values”,*Data Mining and Knowledge Discovery 2*, (1998), pp.283–304.
6. Huang, Z.and Ng, M., “A Fuzzy *k*-Modes Algorithm for Clustering Categorical Data”,*IEEE Transactions on Fuzzy Systems, Vol. 7*, No. 4, August 1999.
7. MacQueen, J. B.,“Some Methods for classification and Analysis of Multivariate Observations”, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, (1967), pp.281–297.
8. Ruspini and Enrique, H., "A new approach to clustering", *Information and control, Vol. 15.1*, (1969), pp.22-32.
9. Zadeh, L.A.,"Fuzzy sets",*Information and Control , Vol.* 8(3),(1965) , pp.338–353.

Corresponding Author:
Akarsh Goyal*,
Email: *akarsh.goyal15@gmail.com*