ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Comparative Analysis of Rough Intuitionistic Fuzzy K-Mode Algorithm for Clustering Categorical Data

Tripathy BK, Akarsh Goyal*, and Patra Anupam Sourav.

School of Computer Science and Engineering, VIT University,Vellore-632014, Tamil Nadu, India

ABSTRACT

In this paper we introduce the concept of rough intuitionistic fuzzy k-mode algorithm to cluster categorical data. This proposal is an extension of rough fuzzy k-mode in which we have added the parameter intuitionistic degree in the calculation of membership values of all elements in a given cluster. The efficiency of the proposed algorithm is demonstrated using various popular categorical data sets from UCI data repository. Experimental analysis is performed by taking these data sets and several measures of efficiency like the DB index, D index, XB index, PC pair and Minkowski score are computed for each data set. The results invariably show that the proposed algorithm is more efficient than rough fuzzy k-mode algorithm.

Keywords - Categorical data, Clustering, Data mining, rough fuzzy k-mode, rough intuitionistic fuzzy k-mode

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Data Mining is the computational process of finding patterns in large data sets involving various methods. These methods are at the intersection of artificial intelligence, machine learning, statistics, and database systems. The data mining process seeks to extract information from a data set and convert this into a comprehendible form for further use. For the extraction of valid patterns and knowledge mining from complex and huge amount of data set many techniques are used. Some of them are association, classification, clustering, pattern recognition etc. These are used to group, or classify the dataset. In this paper we focus on clustering algorithm for mining purposes.

Clustering [8] aims to group a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Hence a cluster is defined as a group of objects which are "similar" to each other 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 database 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.

Both numerical and categorical data can be clustered. 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 clustering algorithms which involve computation of the mean of clusters as a parameter are rendered ineffective when applied on categorical data. This is because they wholly depends on the distance metric and it can only minimize a numerical cost function. So for categorical data we have to use mode type [5] methods.

The mode type approach modifies the means process for clustering categorical data by substituting the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centres 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.

We have used the concept of roughness in this paper. There prevails uncertainty in crisp labelling of data due to context dependent nature. Also, it may be difficult to differentiate distinct objects, and so one may find it convenient to consider granules for its handling. Granulation is a computing paradigm that is abstracted from natural phenomena. The structure of granulation can often be defined by employing various soft computing approaches like rough sets, fuzzy sets [16] or their combination. As a result, rough set approaches have become preferred choices across various application domains, for performing different computing tasks involving vagueness. Rough set theory [11] was basically developed to deal with vagueness in the data. While fuzzy sets deal with such data using a partial membership function, rough sets express the same by the boundary region of a set. A rough set is a set of objects which cannot be classified with certainty as members of the set or its complement using the available knowledge. Thus, associated with every rough set, there is a pair of precise sets known as lower approximation and upper approximation of the rough set. The basic idea is to separate discernible objects from indiscernible ones and to assign them to lower and upper approximations of the set respectively. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability distributions in statistics or a grade of membership in fuzzy set theory [16].

The basic rough k-mode was introduced in [12]. When the notion of fuzziness is applied to rough k-mode we get the rough fuzzy k-mode algorithm. An algorithm to this effect was proposed and studied in [14]. The concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters formed by using methods like k-mode. In fuzzy clustering [2][13] data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the

association between that data element and a particular cluster. Hence it is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

It is well known that the concept of intuitionistic fuzzy sets was introduced by Atanassov [1] is a more general concept than that of fuzzy sets. The presence of the hesitation function controls and presents the uncertainty in a better way. So in this paper we will be working on the idea of rough intuitionistic fuzzy k-mode and also validate that it is better as compared to rough fuzzy k-mode.

RELATED WORK

A lot of work has been done in the field of data mining and data clustering. New methods have been proposed frequently as there is no fixed method of clustering data. Let us first present the history of clustering. In [8] 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 [16]. Applying this concept on clustering Ruspini first proposed the fuzzy clustering algorithm mentioned in [13], which was later modified and generalized by Dunn and Bezdek respectively in [2]. In 1982 Pawlak came out with the concept of rough sets in [11]. The notion of rough k-means clustering was developed by Lingras in [7]. The concepts of k-mode and fuzzy k-mode were introduced by Z. Huang in [5] and [6] respectively. In [9] and [10], Maji and Mitra proposed the notion of rough-fuzzy hybrid clustering algorithm respectively. Ranga Suri and Murty formulated the clustering of data using rough k-mode algorithm in [12]. In [14] the rough fuzzy k-mode algorithm was discussed. Chaira T. formulated the intuitionistic fuzzy clustering algorithm [3], [4]. In [15] the intuitionistic fuzzy k-mode algorithm was presented and studied. The details of all the algorithms have been discussed in the forthcoming sections of the document.

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, wine and iris. The description for these datasets is given in the table below –

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

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.

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

ISSN: 0975-8585

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 ∈ 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 $[x_1, x_2, x_3, \ldots, x_m]$. 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 ≤ j ≤ 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 $A_1; A_2; \ldots; A_m$.

Rough Set

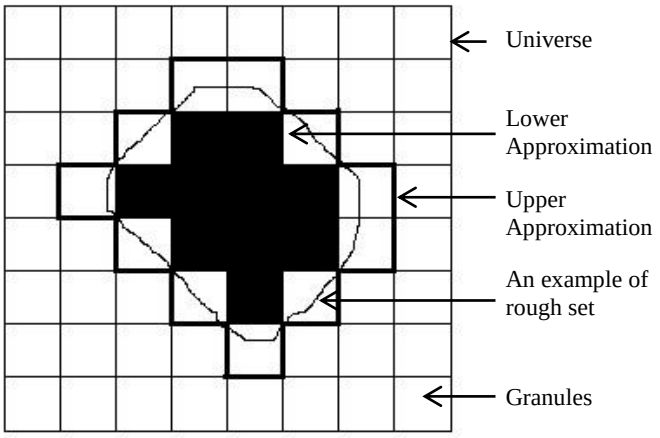

Fig 1: Lower and Upper Approximations of Rough set

Rough set [11] was first described by a Polish computer scientist Pawlak. It is a formal approximation of a crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set. Let X,a subset of U, be a target set that we wish to represent using attribute subset P. The statement that an arbitrary set of objects X comprises a single class, and we wish to express this class using the equivalence classes induced by attribute subset P. In general, X cannot be expressed exactly, because the set may include and exclude objects which are indistinguishable on the basis of attributes P.

For example, consider the target set X={O_1, O_2, O_3, O_4}, and let attribute subset P= {P_1, P_2, P_3, P_4, P_5}, the full available set of features. It will be noted that the set X cannot be expressed exactly, because in $[x]_P$, objects {O_3, O_7, O_{10}} are indiscernible. Thus, there is no way to represent any set X which includes O_3 but excludes objects O_7 and O_{10}.

However, the target set X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of X:

$P X = \{x \mid [x]_P \subseteq X\}$ (1)

$P'X = \{x \mid [x]_P \cap X \neq \phi\}$ (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

ISSN: 0975-8585

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

$$\pi_A(x) = 1 - m_A(x) - n_A(x), \quad 0 \leq \pi_A(x) \leq 1 \qquad (3)$$

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

$$N(m(x)) = (1 - m(x)) / (1 + \lambda m(x)) \quad \lambda > 0, \; N(1) = 0, \; N(0) = 1 \qquad (4)$$

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

$$\pi_A(x) = 1 - m_A(x) - (1 - m_A(x)) / (1 + \lambda m_A(x)). \qquad (5)$$

METHODS AND ALGORITHMS

Rough K-modes algorithm

The membership of the given objects in their clusters is determined by using k-modes algorithm [5]. According to it the objects were repeatedly assigned to different clusters and the cluster centres were also determined in the clusters. By taking into account rough feature with this we take care of the boundary values and also clearly define what is vague and the uncertainty.

Given a categorical data set D with n objects, the objective is to produce k rough clusters $\{U_1, U_2, \ldots, U_k\}$ represented by their modes $\{Z_1, Z_2, \ldots, Z_k\}$ respectively. Let D= { X_1 , X_2 ,... X_n } be the input data set consisting of n data objects, described using m categorical attributes. Each data object Xi is represented as an m-dimensional vector { x_{i1} , x_{i2} ,. . . , x_{im} }.Let freq($x_{i,r}$) denote the number of objects in D with the value $x_{i,r}$ for the rth attribute.

Similarly, let $freq_j^{low}(x_{i,r})$ and $freq_j^{up}(x_{i,r})$ denote the number of objects in the lower and upper approximations of jth cluster respectively with the value $x_{i,r}$ for the rth attribute. Let d(Z_j,X_i) be the distance between a categorical data object X_i and a cluster C_j (with its mode Z_j) during the clustering process. Then, a possible way to compute this distance value between two object X and Y is by using the dissimilarity measure is given below:

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \qquad (6)$$

where

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

ISSN: 0975-8585

The lower and upper approximations are weighted differently. Since the objects in the lower approximation completely belong to the cluster, therefore they are assigned a greater weight denoted by w_{low}. The objects in the upper approximation as assigned a relatively lower weight denoted by w_{up} where $w_{low} + w_{up} = 1$. The algorithm for is given as follows:

1. Assign initial mode Z_i for c clusters.
2. Let $d_{i,k}$ be the minimum and $d_{j,k}$ be the next to minimum distance of x_k from clusters U_i and U_j. Assign each data object to the lower or upper approximation by computing $d_{j,k} - d_{i,k}$.
3. If $d_{j,k} - d_{i,k}$ is less than threshold (ε) then
   $x_k \in \overline{B}U_i$ and $x_k \in \overline{B}U_j$ and is not the member of any lower approximation.
   else $x_k \in \underline{B}U_i$
4. Count cluster-wise attribute value frequencies $freq_j^{low}(x_{i,r})$ and $freq_j^{low}(x_{i,r})$, for every $x_{i,r}$.
5. Calculate new centroids for each cluster using.

$$\begin{cases} \sum_{r=1}^{m}\left(\dfrac{freq_j^{low}(x_{i,r})}{\left|\underline{B}U_i\right|}\right), & if\ \underline{B}U_i \neq \phi\ and\ \left(\overline{B}U_i - \underline{B}U_i\right) = \phi; \\ \sum_{r=1}^{m}\left(\dfrac{freq_j^{up}(x_{i,r})}{\left|\overline{B}U_i\right|}\right), & if\ \underline{B}U_i = \phi\ and\ \overline{B}U_i \neq \phi; \\ \sum_{r=1}^{m}\left(w_{low}\dfrac{freq_j^{low}(x_{i,r})}{\left|\underline{B}U_i\right|} + w_{up}\dfrac{freq_j^{up}(x_{i,r}) - freq_j^{low}(x_{i,r})}{\left|\overline{B}U_i - \underline{B}U_i\right|}\right), & else. \end{cases}$$

(7)

6. Repeat from step 2 until there are no more assignment

Rough Fuzzy K-modes algorithm

Rough Fuzzy K-Mode [14] is an algorithm proposed by Saha, Maulik and Sarkar; it combines the concepts of rough set theory [11] and fuzzy set theory [16]. It is an extension of rough fuzzy C-means [9][10]. The concepts of lower and upper approximations in rough set deals with uncertainty, vagueness and incompleteness whereas the concept of membership function in fuzzy set helps in enhancing and evaluating overlapping clusters.

1. Assign initial mode v_i for c clusters.
2. Compute μ_{ik} using

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}$$

(8)

3. Let μ_{ik} and μ_{jk} be the maximum and next to maximum membership values of object x_k to cluster centroids v_i and v_j.

   If $\mu_{ik} - \mu_{jk} < \varepsilon$ then

$x_k \in \overline{B}U_i$ and $x_k \in \overline{B}U_j$ and x_k cannot be a member of any lower approximation.

Else

$x_k \in \underline{B}U_i$

4. Calculate new cluster mode by using

The mode is updated in such a way that $v_{i,j} = a^r_j \in DOM(A_j)$.

$$r = \arg\max_{1 \le t \le q_j} \begin{cases} w_{low} \, X \, L_{low} + w_{up} \, X \, L_{up}, & if \ \underline{B}U_i \neq \phi \ and \ \overline{B}U_i \neq \phi; \\ L_{low}, & if \ \underline{B}U_i \neq \phi \ and \ \overline{B}U_i = \phi; \\ L_{up}, & else. \end{cases} \quad (9)$$

where,

$$L_{low} = \sum_{\substack{1 \le i \le n, \\ x_i \in \underline{B}U_i, \\ x_{i,j} = a_j^t}} (\mu_{li})^{\eta}, \text{ and } \quad L_{up} = \sum_{\substack{1 \le i \le n, \\ x_i \in \overline{B}U_i, \\ x_{i,j} = a_j^t}} (\mu_{li})^{\eta}$$

5. Repeat from step 2 until termination condition is met or until there are no more assignment of objects.

Rough Intuitionistic Fuzzy K-modes algorithm

We propose a new k-mode algorithm that brings together all the concepts that have been discussed earlier. The rough intuitionistic fuzzy k-mode (RIFKM) uses the concept of rough sets, fuzzy sets and as well as intuitionistic fuzzy sets, thereby making it a perfect combination of IFKM [15] and RKM [12]. It can also be considered to be RFKM [14] with IFS[1], hence adding the concept of lower and upper approximation of rough set, fuzzy membership of fuzzy set, non-membership and hesitation value of intuitionistic fuzzy set. It provides a holistic and all-round approach to clustering of data as it deals with uncertainty, vagueness, incompleteness which, enables the efficient handling of overlapping partitions and improves accuracy.

In RIFKM, each cluster can be identified by three properties, a centroid, a crisp lower approximation and an intuitionistic fuzzy boundary. If an object belongs in the lower approximation of a cluster then its corresponding membership value is 1 and hesitation value is 0. The objects in the lower region have same influence on the corresponding cluster. If an object belongs in the boundary of one cluster then it possibly belongs to that cluster and potentially belongs to another cluster. Hence the objects in the boundary region have different influence on the cluster. Thus we can say that in RIFKM the membership value of objects in lower region is unity $(\mu'_{ij} = 1)$ and for those in boundary region behave like IFKM [15].

The steps that are to be followed in this algorithm are as given below-

1. Assign initial mode v_i for c clusters by choosing any random c objects as cluster.
2. Calculate d_{ik} using equation (6).
3. Compute U matrix
   If $d_{ik} = 0 \ or \ x_j \in \underline{B}U_i$ then
   $\mu_{ik} = 1$
   Else compute μ_{ik} using (8).
4. Compute π_{ik}

ISSN: 0975-8585

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

5. Compute μ'_{ik} and normalize
$$\mu'_{ik} = \mu_{ik} + \pi_{ik} \quad (11)$$

6. Let μ'_{ik} and μ'_{jk} be the maximum and next to maximum membership values of object x_k to cluster centroids v_i and v_j.
If $\mu'_{ik} - \mu'_{jk} < \varepsilon$ then
$x_k \in \overline{B}U_i$ and $x_k \in \overline{B}U_j$ and x_k cannot be a member of any lower approximation.
Else
$x_k \in \underline{B}U_i$

7. Calculate new cluster means by using (9) and substituting u_{li} by $u_{li'}$.

8. Repeat from step 2 until termination condition is met or until there are no more assignment of objects.

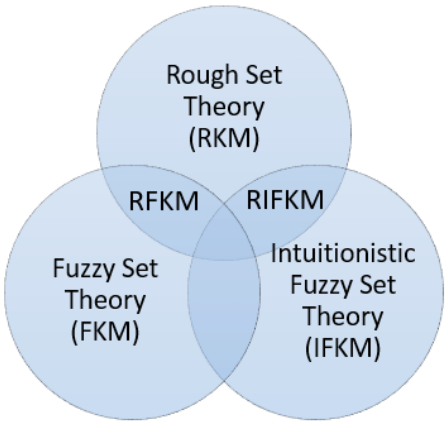

Fig 2: Venn diagram showing combination of theories

1. MEASURING INDICES

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. Also we have calculated the overall accuracy of clustering. The results are dependent on the number of clusters one requires.

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.

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

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.

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

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

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.

Minkowski Score

Minkowski score for a set of n elements is a clustering solution. It can be represented by an n × n matrix C, where $C_{i,j}$ = 1 or 0 depending on whether point i and j are in the same cluster according to the solution. The Minkowski score (MS) where clustering result is C with reference to T, which is the matrix corresponding to the true clustering, is defined as

$$MS(T, C) = \frac{\|T - C\|}{\|T\|} \quad (14)$$

where $\|T\| = \sqrt{\sum_i \sum_j T_{i,j}}$. (15)

Between the two matrices the Minkowski score is the normalized distance. Lower Minkowski score implies better clustering solution, and a perfect solution will have a score zero.

Percentage of Correct Pair

Percentage of Correct Pair (%CP) is given as below

CP = number of pairs correctly clustered into the same cluster/ pairs actually in the same cluster (16)

Higher value of CP signifies the better clustering result. It gives the result in percentage form. Therefore, 100% means perfect clustering.

Xie-Beni Index

Xie-Beni (XB) index is the ratio of the total fuzzy cluster variance (σ) to the minimum separation (ζ) of the clusters. XB can be defined as below.

XB = $\frac{\sigma}{n \, X \, \zeta}$ Here total number of data objects is n (17)

Where

$$\sigma = \sum_{l=1}^{k} \sum_{i=1}^{n} (\mu_{li})^2 D(v_{l,} x_i) \quad (18)$$

and

$$\zeta = \min_{h \neq l} \{ D(v_{h,} v_l) \} \quad (19)$$

The dissimilarity measure between cluster mode v_l and object x_i is $D(v_l, x_i)$.
Lower value of XB gives better clustering result.

ISSN: 0975-8585

Clustering Accuracy

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

$$r = \frac{\sum_{l=1}^{k} a_l}{n} \qquad (20)$$

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.

RESULTS AND ANALYSIS

To evaluate the performance and efficiency of the rough intuitionistic fuzzy k-modes algorithm and compare it with the rough fuzzy k-modes algorithm 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 rough intuitionistic fuzzy k-modes algorithm we specified λ= 2. w_{low} and w_{up} are assigned the values 0.7 and 0.3 respectively. $\in$ is taken as 1.1.

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 rough intuitionistic fuzzy k-modes and rough fuzzy k-modes algorithms indeed produce different clusters.

Modes of the Clusters

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

Soybean dataset

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

Table 2: Columns 1 through 22 for Rough 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	6	0	2	0	0	3	0	1	0	2	1	0	0	2	2	0	0	0	1	0	3	1
2	6	0	0	1	1	3	3	1	1	0	1	1	0	2	2	0	0	0	1	0	0	3
3	1	1	0	0	1	0	3	2	1	0	1	0	0	2	2	0	0	0	1	0	0	2
4	0	1	1	1	0	2	1	2	0	1	1	0	0	2	2	0	0	0	1	1	1	1

Table 3: Columns 23 through 36 for Rough fuzzy K-Mode

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

ISSN: 0975-8585

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

Table 4: Columns 1 through 22 for Rough 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	5	0	2	1	1	3	0	2	1	0	1	1	0	2	2	0	0	0	1	0	3	1
2	4	0	0	1	0	1	2	1	0	1	1	1	0	2	2	0	0	0	1	1	0	3
3	3	1	2	0	0	0	3	1	0	2	1	1	0	2	2	0	0	0	1	0	1	2
4	1	0	2	0	1	0	1	1	0	0	1	0	0	2	2	0	0	0	1	1	1	1

1.

Table 5: Columns 23 through 36 for Rough 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	0	0	0	2	1	0	4	0	0	0	0	0	0	2
3	0	0	0	0	0	3	4	0	0	0	0	0	1	4
4	0	1	1	0	0	3	4	0	0	0	0	0	0	3

Iris Dataset

Table 6 shows the results obtained by using the Rough fuzzy k-mode algorithm and Rough intuitionistic fuzzy k-mode algorithm.

Table 6: Columns 1 through 4

	Rough Fuzzy K-mode				Rough Intuitionistic Fuzzy K-mode			
Z_i	1	2	3	4	1	2	3	4
1	6.3	2.5	5.7	1.9	6.7	3	5.2	2.5
2	6.7	2.5	5	2.5	5.8	3.3	5.7	2.3
3	6.3	2.5	5	1.9	6.7	3.1	5.7	2.2

Wine Dataset

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

Table 7: Columns 1 through 13 for Rough fuzzy K-mode algorithm

	Rough Intuitionistic Fuzzy K-mode				
Datasets	DB	D	Minkowski Score	% of Correct Pair	Xie-Beni Index
Soybean	1.0348	0.9231	0.2355	94.57	0.507
Iris	2.6394	0.75	0.3992	67.97	0.71
Wine	9.2695	0.1538	0.25	84.65	2.6233

Table 8: Columns 1 through 13 for Rough intuitionistic Fuzzy K-mode algorithm

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13
1	14.13	4.1	2.74	20	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560
2	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
3	14.13	4.1	2.74	24.5	120.	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560

Performance Metrics Values

Now we have calculated the DB-index, D-index, Minkowski Score, Percentage of Correct Pair and Xie-Beni Index of the two algorithms on the three datasets. The representation for this has been made with the help of a table shown below and bar-graphs which clearly indicate that rough intuitionistic fuzzy k-mode is better than rough fuzzy k-mode.

ISSN: 0975-8585

Table 9: Performance Metrics Values for Rough Fuzzy K-mode

	Rough Fuzzy K-mode				
Datasets	DB	D	Minkowski Score	% of Correct Pair	Xie-Beni Index
Soybean	2.054	0.55	0.4879	71.14	1.3104
Iris	6.5467	0.25	0.4817	66.05	2.206
Wine	9.3477	0.1538	0.3997	69.61	2.9849

Table 10: Performance Metrics Values for Rough Intuitionistic Fuzzy K-mode

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13
1	14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.06	9.2	0.61	1.6	560
2	13.27	4.28	2.26	20.5	95	1.59	0.69	0.52	1.35	10.2	0.59	1.74	835
3	14.13	4.1	2.74	24.5	105	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560

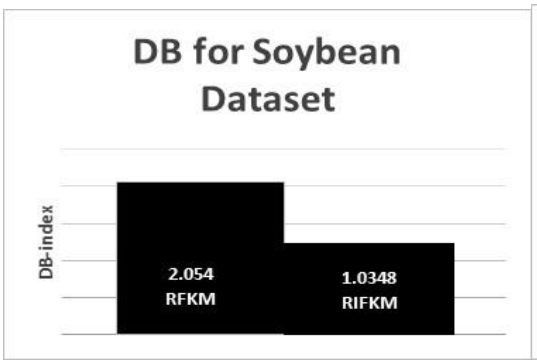

Fig 3: Graph of DB for Soybean Dataset

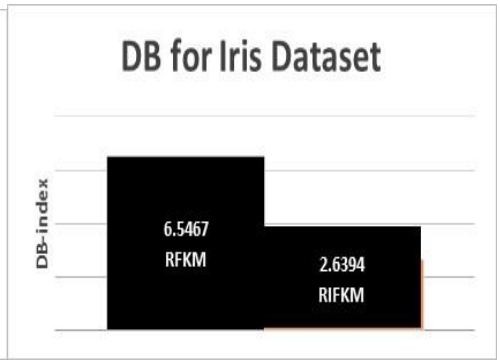

Fig 4: Graph of DB for Iris Dataset

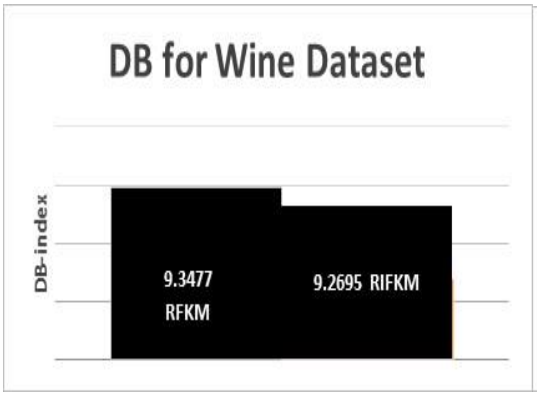

Fig 5: Graph of DB for Wine Dataset

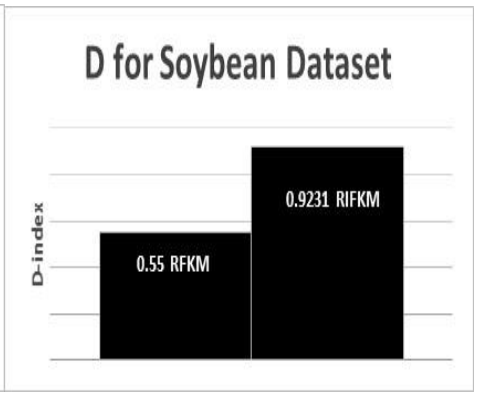

Fig.6: Graph of D for Soybean Dataset

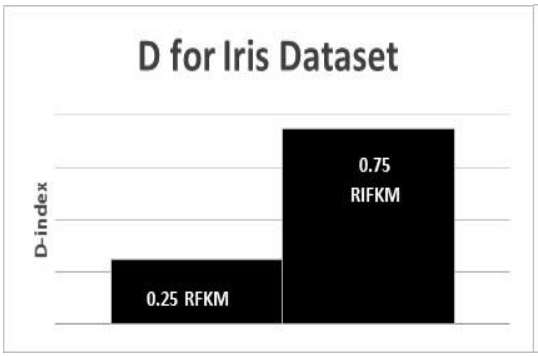

Fig.7: Graph of D for Iris Dataset

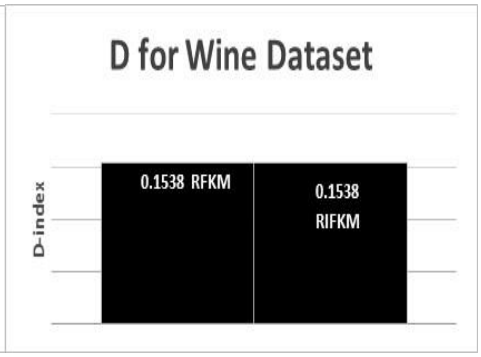

Fig.8: Graph of D for Wine Dataset

ISSN: 0975-8585

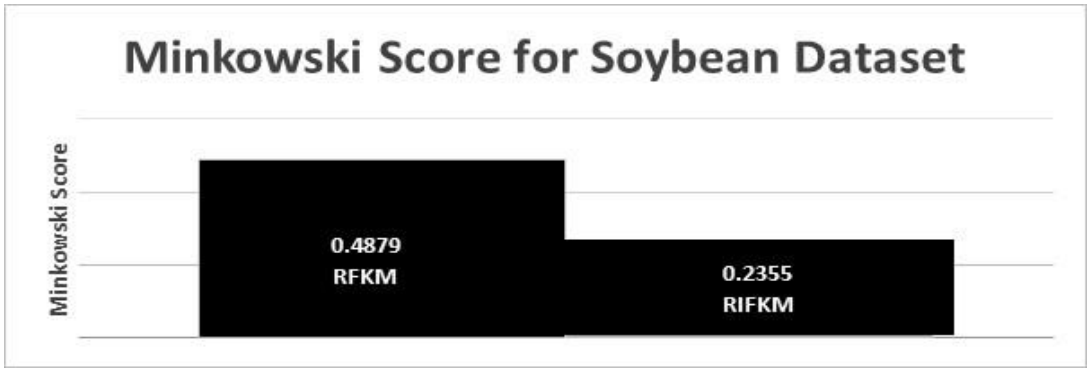

Fig.9: Graph of Minkowski Score for Soybean Dataset

Fig.10: Graph of Minkowski Score for Iris Dataset

Fig.11: Graph of Minkowski Score for Wine Dataset

Fig.12: Graph of % of Correct Pair Soybean Dataset

ISSN: 0975-8585

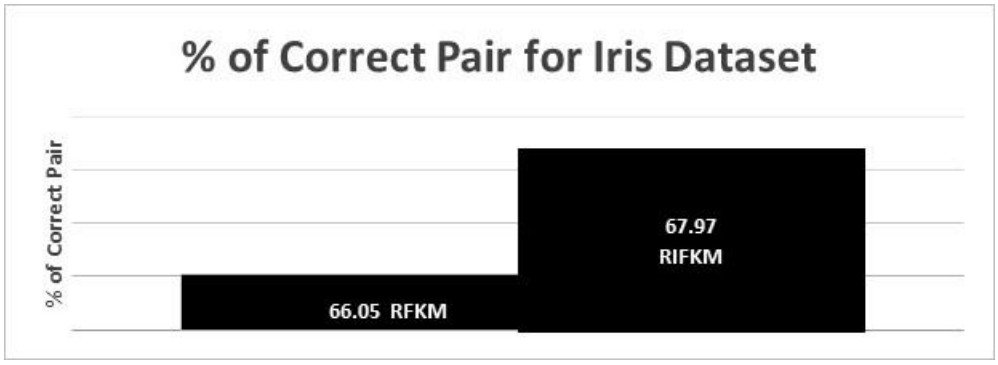

Fig.13: Graph of % of Correct Pair Iris Dataset

Fig.14: Graph of % of Correct Pair Wine Dataset

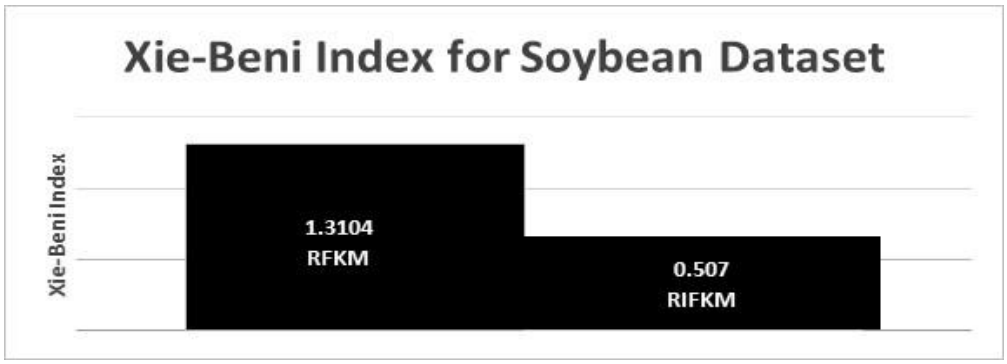

Fig.15: Graph of Xie-Beni Index Soybean Dataset

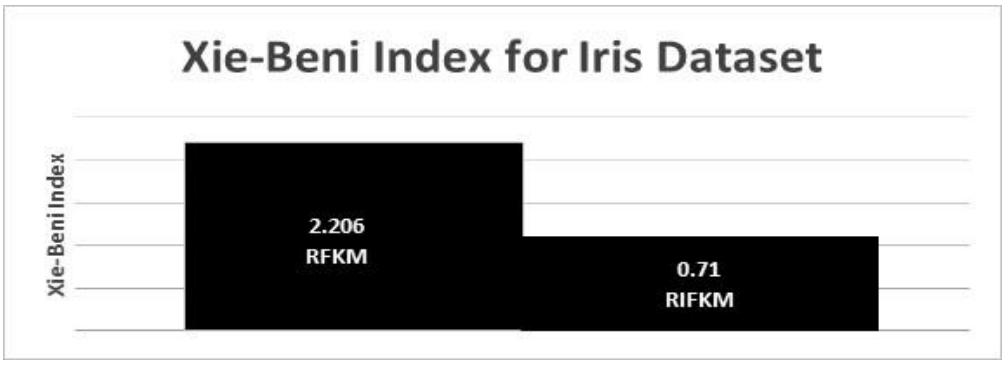

Fig.16: Graph of Xie-Beni Index Iris Dataset

ISSN: 0975-8585

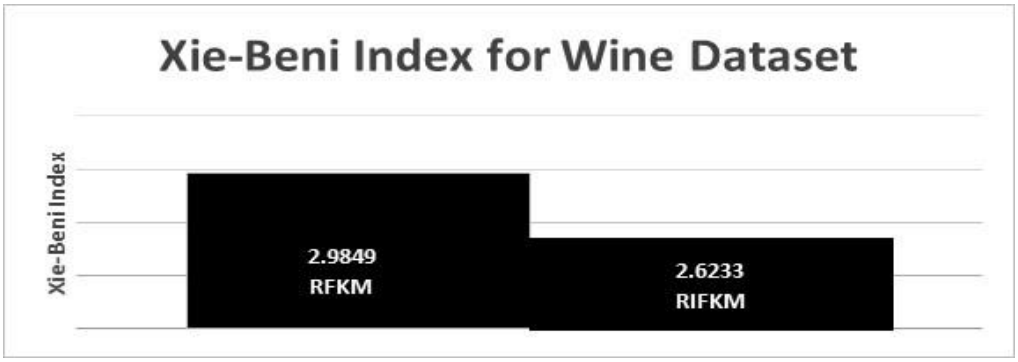

Fig.17: Graph of Xie-Beni Index Iris Dataset

Accuracy

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

Table 11: Accuracy of Clustering

Datasets	Rough Fuzzy K-mode	Rough Intuitionistic Fuzzy K-mode
Soybean	0.3325	0.748
Iris	0.607	0.723
Wine	0.58	0.637

According to Table 11 the final outcome is true. The accuracy results obtained clearly justify that rough intuitionistic fuzzy k-mode is a much better algorithm for clustering categorical data than rough fuzzy k-mode.

CONCLUSION

Categorical data have become necessary in the real-world databases. However, few efficient algorithms are available for clustering massive categorical data. Earlier k-mode and fuzzy k-mode algorithms were taken up to tackle these problems. But the issue of uncertainty and vagueness were not being addressed by these notions. The uncertainty prevailing in clustering is addressed by considering a soft computing approach based on rough sets. A novel rough clustering algorithm was designed by modifying the standard k-modes algorithm to incorporate rough sets principles which was later extended to incorporate the membership function which led to the development of rough fuzzy k-mode algorithm. So in this paper we proposed the rough 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. Several new measures are defined based on rough sets to evaluate the performance of rough–fuzzy clustering algorithms. The superior performance of the proposed rough intuitionistic fuzzy k-modes algorithm is experimentally established on various benchmark categorical data sets. 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. Hence, the rough intuitionistic fuzzy k-mode has proven to be an important enrichment to clustering approaches, particularly in the direction of soft computing methods.

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 different threshold values provides different set of cluster. So according to our application it can be changed for better result. The notion discussed here could be applied for the detection of any outliers as well. Also, one can work on establishing a formal relationship among various parameters of the proposed algorithm.

ISSN: 0975-8585

REFERENCES

[1] Atanassov, K., "Intuitionistic Fuzzy Sets", Fuzzy Sets and Systems, 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] Lingras, P. and West, P., "Interval set clustering of web users and rough k-means", Dept. math Comput. Sci., St. Mary's Univ., Halifax, NS, Canada, Tech. Rep. No. 2002-002, 2002.

[8] 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.

[9] Maji, P., and S.K. Pal, "RFCM: A Hybrid Clustering Algorithm using rough and fuzzy sets", Fundamental Informaticae, 80.4, (2007): 475-496.

[10] Mitra, S. and Banka, H. and Pedrycz, W., "Rough-Fuzzy Collaborative Clustering", IEEE Transactions on systems, man and cybernetics-Part B: Cybernetics, Vol. 36 No. 4, August 2006.

[11] Pawlak and Zdzisław , "Rough sets", International Journal of Parallel Programming ,Vol.11 (5), (1982), pp. 341–356.

[12] Ranga Suri ,N. N. R. and Murty, " Detectitng outliers in categorical data through rough clustering", Nat Comput,(2015), DOI 10.1007/s11047-015-9489-2.

[13] Ruspini and Enrique, H., "A new approach to clustering", Information and control, Vol.15.1, (1969), pp.22-32.

[14] Saha, I. and Sarkar, J. and Maulik, U. , "Rough Set Based Fuzzy K-Modes for Categorical Data", Swarm, Evolutionary, and Memetic Computing: Third International Conference, SEMCCO 2012, Bhubaneswar, India, December 20-22, 2012, pp.323-330

[15] Tripathy, B.K. and Goyal, A. and Patra, A.S., "Clustering Categorical Data Using Intuitionistic Fuzzy K-mode", Communicated to International Journal of Pharmacy and Technology, (2016).

[16] Zadeh, L.A., "Fuzzy sets", Information and Control, Vol.8(3), (1965) , pp.338–353