International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 9 (2017) pp. 232-240

Received: 2 December 2016; Accept: 21 August 2017; Publish: 7 September 2017

A Comparative Analysis of Simulated Annealing Based Intuitionistic Fuzzy K-Mode Algorithm for Clustering Categorical Data

Akarsh Goyal[1], Patra Anupam Sourav[2] and Arunkumar Thangavelu[3]

[1] School of Computer Science and Engineering, VIT University,
Vellore, India
akarsh.goyal15@gmail.com

[2] School of Computer Science and Engineering, VIT University,
Vellore, India
anupam.sourav@gmail.com

[3] School of Computer Science and Engineering, VIT University,
Vellore, India
arunkumart@vit.ac.in

***Abstract*: In this paper we introduce the concept of simulated annealing on intuitionistic fuzzy k-mode algorithm to cluster categorical data. This notion is an extension of intuitionistic fuzzy k-mode in which we have added the concept of energy related objective functions, temperature ranges and probability so as to provide better clusters for the data objects. There is a deep and useful connection between statistical mechanics and the kind of multivariate optimization we are doing here. A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods. So simulated annealing has been used here. Also the intuitionistic parameter has been retained for 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 intuitionistic fuzzy k-mode. Simulated Annealing based Intuitionistic fuzzy k-mode is very efficient when clustering large categorical data sets, which is very much critical to data mining applications.**

***Keywords*:** Categorical data, Clustering, Data mining, Probability Intuitionistic fuzzy k-mode, Simulated Annealing

I. Introduction

Data Mining is the computational process of finding patterns in large data sets involving methods 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 a comprehendible 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 such a way that objects in the same group are more similar to each other than to those in other groups [1]. Hence a *cluster* is a group of objects which are "similar" between themselves and are "dissimilar" to the objects belonging to other clusters. Most of the raw data available nowadays 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. These methods 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 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 1 below.

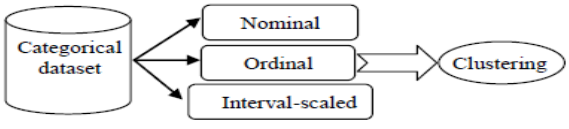

Figure 1. Clustering Categorical Data

Clustering can be performed for both numerical and categorical data. But clustering categorical data is very different and difficult from those of numerical data. There can't be any direct application of the distance metric to the categorical data directly. So k-means algorithm which is the most used clustering method is rendered inefficient when applied on categorical data. This is because it fully depends on the distance metric and it can only minimize a numerical cost function. So for categorical data we have to use methods which deal with finding the modes.

The k-modes [2] 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 [3] 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 [4]. 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 k-modes algorithm by assigning confidence to objects in different clusters. Also, it is well known that the concept of intuitionistic fuzzy set [5] was developed by Atanassov, in which it was indicated that there exists an intuitionistic degree, $\pi_A(x)$, which arises due to lack of knowledge in defining membership degree. Using this an addendum of fuzzy k-modes known as intuitionistic fuzzy k-modes was presented.

In this paper we have devised a new model based on the concepts stated above called simulated annealing based intuitionistic fuzzy k-modes. In this notion we have used the properties of simulated annealing [6] on intuitionistic fuzzy k-mode. This has been done so as to get better outcome than intuitionistic fuzzy k-mode while clustering [7] categorical data.

II. Related Work

An immense amount of work 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. The major papers related to clustering have been expounded below -

In [1] 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 [8]. Applying this concept on clustering Ruspini first proposed the fuzzy clustering algorithm mentioned in [4], which was later modified and generalized by Dunn and Bezdek respectively in [9]. The concepts of k-mode and fuzzy k-mode were introduced by Z. Huang in [2] and [3] respectively. In 1986 Atanassov K T. developed the intuitionistic fuzzy set theory written in [5], on the basis of which Chaira T. formulated the intuitionistic fuzzy clustering algorithm [10], [7]. Kirkpatrick et al introduced the concept of simulated annealing for optimization in [6]. Saha et al gave the simulated annealing method for clustering categorical data using K-medoid in [11]. In [12] fuzzy centroid clustering method has been proposed. Tripathy et al introduced the intuitionistic fuzzy k-mode method in [13] and expanded it using roughness theory in [14]. The details of all the algorithms have been discussed in the forthcoming sections of the document.

III. 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 glass dataset, iris dataset and wine dataset. Glass dataset has 1 to 7 class values and the number of instances are 214. The number of attributes contained by it are 10. Whereas the wine dataset has 178 objects and 13 attributes. Also there are only 3 classes in it. Finally the iris dataset has 3 classes. It has 150 objects and 4 features. All these datasets are multivariate and the main task associated with these are that of classification.

IV. 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 simulated annealing based intuitionistic fuzzy k-mode have been provided.

A. 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 $[x_1, x_2, x_3, \ldots, x_m]$. 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 $X = \{X_1, X_2, \ldots X_n\}$ be a set of n objects. Object X_i is represented as $[x_{i1}, x_{i2}, \ldots, x_{im}]$. 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 $A_1; A_2; \ldots; A_m$.

B. Intuitionistic Fuzzy Set

The notion of intuitionistic fuzzy sets introduced by Atanassov [5] 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 $\{(x, m_A(x), n_A(x)) \mid x \in X\}$, where $m_A : X \to [0,1]$ and $n_A : X \to [0,1]$ such that

$0 \le m_A(x) + n_A(x) \le 1$ 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

$$\pi_A(x) = 1 - m_A(x) - n_A(x),\ 0 \le \pi_A(x) \le 1 \quad (1)$$

Membership values $m_A(x)$ lie in an interval range $\left[m_A(x) - \pi_A(x),\ m_A(x) + \pi_A(x)\right]$ 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

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

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

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

V. Methods and Algorithms

A. 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

$$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j; \\ 1, & x_j \ne y_j. \end{cases} \quad (5)$$

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.

B. K-modes algorithm (KM)

In general, k-modes clustering [2] 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:

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \omega_{li}\ d(Z_l, X_i) \quad (6)$$

subject to $\omega_{li} \in \{0,1\},\ 1 \le l \le k, 1 \le i \le n$

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

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

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 [12].

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:

$$F(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{i,l} d(x_{ij}, z_{lj}) \quad (8)$$

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 $d(.,.)$.

C. Fuzzy k-modes algorithm (FKM)

The fuzzy k-modes algorithm [3] 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 data in different clusters. In the fuzzy k-modes algorithm, data *D* is grouped into *k* clusters by minimizing the cost function (6-7)

Subject to –

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

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).

D. Intuitionistic Fuzzy K-mode Algorithm (IFKM)

The Intuitionistic fuzzy k-modes [13]-[14] follows 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). Other notations are given in above sections.
4. Compute the hesitation matrix π using
$$\pi_A(x) = 1 - \mu_A(x) - \frac{1-\mu_A(x)}{1+\lambda\mu_A(x)} \mid x \in X \qquad (10)$$
5. Compute the modified membership matrix U' using
$$\mu'_{ik} = \mu_{ik} + \pi_{ik} \qquad (11)$$
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 $\|U'^{(r)} - U'^{(r+1)}\| < \varepsilon$ then stop else repeat from step 4.

The objective function of IFKM contains two terms: (i) modified objective function of conventional FKM using Intuitionistic fuzzy set and (ii) intuitionistic fuzzy entropy (IFE). IFKM minimizes the objective function as:

$$J_{IFKM} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{*m} d_{ik}^2 + \sum_{i=1}^{c} \pi_i^* e^{1-\pi_i^*} \quad \text{where}$$

$$\pi_i^* = \frac{1}{N}\sum_{k=1}^{n} \pi_{ik}, \quad k \in [1, N]. \qquad (12)$$

E. Simulated Annealing Intuitionistic Fuzzy K-mode Algorithm (SAIFKM)

Simulated annealing (SA) [6][11] is a random search technique proposed by Kirkpatrick et al. based on the principles of statistical mechanics regarding the behavior of a large number of atoms at low temperature, for combinatorial optimization problem by optimizing the associated energy. In statistical mechanics, the low energy states of matter are achieved at very low temperatures. A low energy state means a high ordered state. To achieve this, the solid is heated to a very high temperature and then it is decreased slowly until the material freezes into a good crystal ensuring that one spends sufficient time at each temperature value. It is to be noted that the probability of changing a energy state, E_1 to another energy state, E_2 is $e^{\frac{-(E_1-E_2)}{b_T}}$, where b stands for Boltzmann constant and T is the temperature of the system. This connection of annealing in solids provides a full proof method for optimization of features of very large and difficult processes. A great perspective and novel knowledge on mining problems is provided by it.

In order to apply the simulated annealing strategy to a optimization problem, it is necessary to define the objective function (energy of the system), temperature of the system and the schedule (number of annealing iterations). Once all these are defined, SA algorithm follows the basic steps. t is the temperature which is initially set to a maximum value and gradually it is decreased by small value controlled by g. Initial solution, S_{init} is set by random. In each iteration, the solution found so far is perturbed to get new solution. The new solution is assessed through objective function to check whether that solution will be accepted for the next iteration or not. If the perturbed solution is better than the previous one, then it is accepted directly, otherwise it is accepted for the next iteration with the probability $e^{-\left[\frac{((F(S_{per})-F(S_{init}))}{t}\right]}$. The whole process continues until a minimum temperature is reached. So here we have extended intuitionistic fuzzy k-mode [13] by adding the concept of simulated annealing to it.. The procedure is as follows –

Procedure for SAIFKM

Input: X, the dataset η, the fuzzy exponent
ϱ, a small real threshold value between [0,1]
K, the number of cluster
ω_{low}, relative weight for *Lower Approximation* of rough clustering, $0 < \omega_{low} < 1$ (T_{max}),
Maximum temperature (T_{min}),
Minimum temperature *maxItr*,
Maximum Iteration g, a small real value between [0, 1], $0 < g < 1$
Output: [μ_{li}] *where*, $1 \le l \le K$ *and* $1 \le i \le n$

1: Select random K objects from dataset for K cluster mode to encode in String, S_{init}.
2: Set $t = T_{max}$.
3: Calculate energy value $F(S_{init})$ using Eqn. (12) for S_{init}.
4: Calculate μ_{li} for all n objects using (7).
5: Classify objects using algorithm given in section 5.4 and update [μ_{li}].
6: Compute new modes using (11).
7: Update each modes in S_{init} with new modes.
8: **repeat**
9: $S_{per} \leftarrow Perturb(S_{init})$.
10: Calculate μ_{li} for all n objects using (7).
11: Cluster objects using algorithm given in section 5.4 and update [μ_{li}].
12: Calculate the energy value $F(S_{per})$ using Eqn. (12) for S_{per}.
13: **if** $(F(S_{per}) - F(S_{init})) < 0$ **then**
14: $Sinit \leftarrow Sper$ and $F(Sinit) \leftarrow F(Sper)$
15: **else**
16: $S_{init} \leftarrow S_{per}$ and $F(S_{init}) \leftarrow F(S_{per})$ with $e^{-\left[\frac{((F(S_{per})-F(S_{init}))}{t}\right]}$ probability
17: **end if**
18: **until** *maxItr* is reached
19: Set $t \leftarrow t \times g$, where $0 < g < 1$
20: **until** $t < T_{min}$

The simulated annealing based intuitionistic fuzzy k-mode starts with the encoding of the string/configuration as K number of cluster modes which are selected randomly from the

dataset, *X*, while ensuring they are distinct. These *K* modes are encoded in a string as row vector. As there are *m* attributes for each object, so the length of the string is *m*×*K*, where first *m* position represents the first mode, next *m* position represent the second mode and so on. Fuzzy membership matrix is generated for each object using (7). Thereafter the object x_i is classified and the fuzzy membership value of x_i is set accordingly. After clustering, the energy function of a string and the modes are updated using (12) and (11), respectively. And then the process continues.

VI. 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.

A. 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.

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{k\neq i}\left\{\frac{S(v_i)+S(v_k)}{d(v_i,v_k)}\right\} \quad \text{for } 1 < i,k < c \tag{13}$$

v_1 and v_k are the different centroids. c is the number of clusters. 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.

B. 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

$$Dunn = \min_i\left\{\min_{k\neq i}\left\{\frac{d(v_i,v_k)}{\max_i S(v_i)}\right\}\right\} \quad \text{for } 1 < k,i,l < c \tag{14}$$

The notation is same as explained in Davis-Bouldin Index section. 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.

C. Clustering Accuracy

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

$$r = \frac{\sum_{l=1}^{k} a_l}{n} \tag{15}$$

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.

VII. Results and Analysis

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

The datasets used were the glass 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 [13]-[14] we specified $\lambda = 2$. The record was assigned X_i was assigned to the Lth cluster if $\omega_{li} = \max_{1\leq h\leq k}\{\omega_{hi}\}$. For simulated annealing based intuitionistic fuzzy k-mode the different values taken for the constants used in the algorithm were (T_{max}) = 100, (T_{min}) = 0.01, g= 0.9 and *maxItr* = 100.

If the maximum was not unique, then Xi was assigned to the cluster of first achieving the maximum.

We have taken 6, 3 and 3 as the number of clusters for glass, 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 simulated annealing based intuitionistic fuzzy k-modes and intuitionistic fuzzy k-modes algorithms indeed produce different clusters.

A. Modes of the Clusters

In this section we compute the cluster centres for intuitionistic fuzzy k-mode and simulated annealing based intuitionistic fuzzy k-modes for three data sets; glass dataset, iris dataset and wine dataset to show the superiority of simulated annealing based intuitionistic fuzzy k-mode algorithm over the intuitionistic fuzzy k-mode algorithm. The different columns are for the attributes present in the dataset. The rows are for the different clusters formed.

1) Glass dataset

Tables 1 and 2 show the results obtained by using the Intuitionistic Fuzzy k-mode algorithm and Simulated Annealing based Intuitionistic Fuzzy k-mode algorithm respectively.

Table 1. Columns 1 through 9 for intuitionistic fuzzy k-mode

Z_i	1	2	3	4	5	6	7	8	9
1	1.5174	12.78	3.69	0.82	70.43	0.31	8.04	0.76	0.21
2	1.5174	12.96	2.96	0.78	72.92	2.7	8.04	0.14	0.21
3	1.5174	12.96	2.96	0.82	72.92	0.31	8.04	0.14	0.03
4	1.5313	10.73	2.96	2.1	69.81	2.7	13.3	3.15	0.03
5	1.5174	12.96	2.96	0.82	72.92	2.7	8.04	0.14	0.03
6	1.5313	10.73	1.78	2.1	69.81	2.7	13.3	3.15	0.21

Table 2. Columns 1 through 9 for simulated annealing based intuitionistic fuzzy k-mode

Z_i	1	2	3	4	5	6	7	8	9
1	1.5168	12.96	2.19	1.66	72.92	0.1	9.32	0.14	0.34
2	1.5168	14.09	2.19	1.66	71.35	0.45	9.32	3.15	0.03
3	1.5168	14.09	2.19	1.66	74.45	0.1	9.32	0.15	0.34
4	1.5168	14.09	2.19	1.66	72.92	2.7	9.32	0.14	0.01
5	1.5168	12.96	2.96	1.98	72.92	0.97	13.3	0.14	0.24
6	1.5169	12.96	3.69	0.65	72.52	0.58	8.04	0.14	0.03

2) Iris Dataset

Table 3 shows the results obtained by using the Simulated Annealing based Fuzzy k-mode algorithm and Intuitionistic Fuzzy k-mode algorithm.

Table 3. Columns 1 through 4 for IFKM and SAIFKM

	Simulated Annealing Based Intuitionistic Fuzzy K-mode				**Intuitionistic Fuzzy K-mode**			
Z_i	1	2	3	4	1	2	3	4
1	6.7	4.4	4.1	0.5	4.3	2	1.1	0.6
2	4.3	2.0	1.0	0.5	7	4.1	1	0.6
3	7.1	2.3	6.7	0.6	7	2	3.6	0.5

3) Wine dataset

Tables 4 and 5 show the results obtained by using the Intuitionistic Fuzzy k-mode algorithm and Simulated Annealing based Intuitionistic Fuzzy k-mode algorithm respectively.

Table 4. 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	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 5. Columns 1 through 13 for Simulated Annealing based Intuitionistic Fuzzy K-mode algorithm

Z_i	1	2	3	4	5	6	7	8	9	10	11	12	13
1	12.82	3.86	2.22	20.4	105	1.1	0.78	0.35	1.04	2.3	0.65	3	406
2	11.03	3.88	1.9	18	81	1.63	0.78	0.13	0.86	8.21	0.65	2.44	855
3	13.29	2.83	2.75	25.5	82	1.89	1.92	0.41	1.99	10.26	1.42	2.83	406

B. 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 simulated annealing based intuitionistic fuzzy k-mode is better than intuitionistic fuzzy k-mode.

Table 6. DB and D-index values

	Intuitionistic Fuzzy K-mode		**Simulated Annealing based Intuitionistic Fuzzy K-mode**	
Datasets	DB	D	DB	D
Glass	11.1	0.1111	4.3869	0.3333
Iris	2.667	0.75	1.7266	1
Wine	2.1667	0.9231	2.15	0.94

Figure 2. Graph of DB for Glass Dataset.

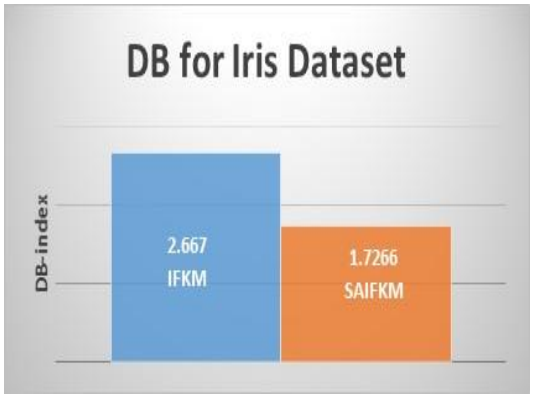

Figure 3. Graph of DB for Iris Dataset.

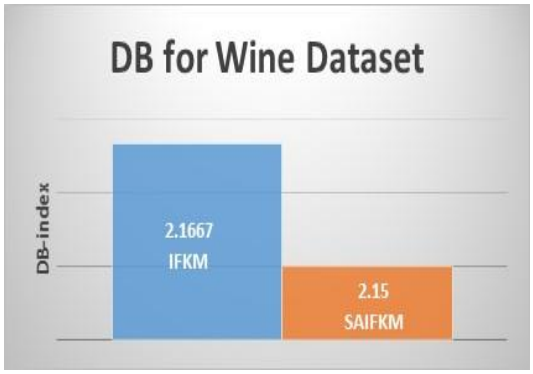

Figure 4. Graph of DB for Wine Dataset.

Figure 5. Graph of D for Glass Dataset

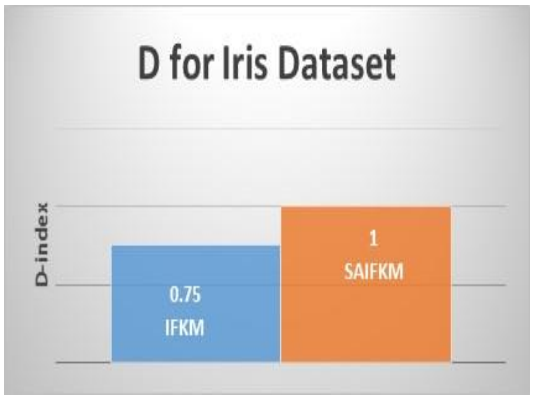

Figure 6. Graph of D for Iris Dataset.

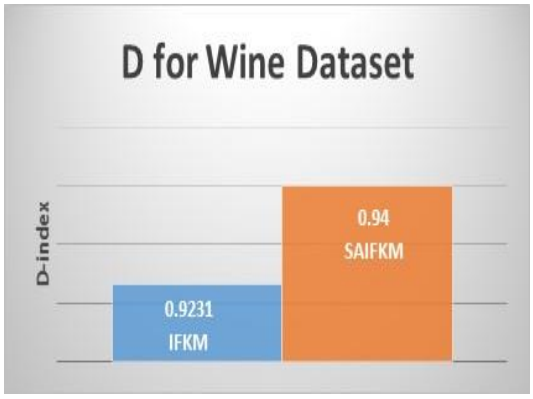

Figure 7. Graph of D for Wine Dataset.

C. *Accuracy*

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

Table 7: Accuracy of Clustering

Datasets	Intuitionistic Fuzzy K-mode	Simulated Annealing based Intuitionistic Fuzzy K-mode
Glass	0.67	0.71
Iris	0.555	0.58
Wine	0.624	0.65

According to Table 7 the final outcome is true. The accuracy results obtained clearly justify that simulated annealing based intuitionistic fuzzy k-mode is a much better algorithm for clustering categorical data than intuitionistic fuzzy k-mode.

VIII. Conclusion

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 k-modes algorithm for clustering categorical objects based on extensions to the fuzzy k–means algorithm was motivated to solve this problem. Later intuitionistic fuzzy k-mode method was proposed in which the intuitionistic degree was taken into effect. This degree led to an uncertainty in the membership of an object in a particular cluster by a particular value. Building upon these methods, we introduced simulated annealing based intuitionistic fuzzy k-mode technique. This process used other parameters which were very much different from those used in intuitionistic fuzzy k-mode. 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. Also the answer found out at the end is a global minima. From the obtained results we perceive that the simulated annealing based intuitionistic fuzzy k-mode algorithm performs better than the intuitionistic fuzzy k-mode algorithm as 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.

IX. 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. In addition to this, the maximum temperature, minimum temperature, factor and number of iterations could be

varied and a concrete relation between these things could help get much better clusters.

References

[1] 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.
[2] 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.
[3] 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.
[4] Ruspini and Enrique, H., "A new approach to clustering", Information and control, 15.1, (1969), pp.22-32.
[5] Atanassov, K., "Intuitionistic Fuzzy Sets", Fuzzy Sets and Systems, 20, (1986), pp.87-96.
[6] Kirkpatrick, S., Gelatt, C. and Vecchi, M. "Optimization by simulated annealing," *Science*, (1983), vol. 220, no. 4598, pp. 671–680.
[7] 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.
[8] Zadeh, L.A., "Fuzzy sets". Information and Control, 8(3), (1965), pp.338–353.
[9] Bezdek, J.C., "Pattern Recognition with Fuzzy Objective Function Algorithms", Kluwer Academic Publishers, (1981).
[10] 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.
[11] Saha, I. and Mukhopadhyay, A., "Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering", Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, (2008), 19-21 March, Vol I.
[12] Kim, D., Lee, K. and Lee, D., "Fuzzy clustering of categorical data using fuzzy centroids", Pattern Recognition Letters, Vol.25 (11), (2004), pp. 1263–1271.
[13] Tripathy, B.K., Goyal, A. and Patra, A.S., "Clustering Categorical Data Using Intuitionistic Fuzzy K-mode", International Journal of Pharmacy and Technology, Vol. 8 (3), September (2016), pp. 16688-16701.
[14] Tripathy, B.K., Goyal, A. and Patra, A.S., " A Comparative Analysis of Rough Intuitionistic Fuzzy K-mode for Clustering Categorical Data", Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol. 7(5), (2016), pp. 2787-2802.

Author Biographies

Akarsh Goyal, He is a student, pursuing a Bachelors of Technology in Computer Science and Engineering at VIT University, Vellore, Tamil Nadu, India. He is an avid reader and computer enthusiast. He has published a few scientific papers in area of IoT, data mining, software engineering and marketing. His areas of interest are data mining, machine learning, intelligent systems, IoT, and android based applications. Akarsh has a passion for design which is reflected by his participations in hackathons. He has a particular aptitude to projects with real life application and scope using computer applications.

Patra Anupam Sourav, Student, School Of Computer Science and Engineering, VIT University, Vellore. He is a regular competitive programmer and has experience in algorithm design and solving complex problems. His areas of interest are data mining, machine learning and big data analytic

Arunkumar Thangavelu is a senior professor in the school of computing sciences and engineering, VIT University, at Vellore, India. He is a researcher and an academician working on Mobile networks, spanning over MANETS, Wireless technology, involving QoS, location management, Handoff Issues and Security. He has published more than 100 technical papers in international journals/ proceedings of international conferences/ edited book chapters of reputed publications like Springer. He has 20 years of teaching experience