DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **17**, No 4
Sofia • 2017 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2017-0044

Application of Genetic Algorithm Based Intuitionistic Fuzzy k-Mode for Clustering Categorical Data

Akarsh Goyal, Patra Anupam Sourav, Kalyanaraman P.

School of Computer Science and Engineering, VIT University, Vellore, India
E-mails: akarsh.goyal15@gmail.com anupam.sourav@gmail.com pkalyanaraman@vit.ac.in

Abstract: *In present times a great number of clustering algorithms are available which group objects having similar features. But most of the datasets have data values that are categorical, which makes it difficult to implement these algorithms. The concept of genetic algorithm on intuitionistic fuzzy k-Mode method is proposed in the paper to cluster categorical data. This model is an extension of intuitionistic fuzzy k-Mode in which the notion of fitness related objective functions, crossovers, mutations and probability has been added to provide better clusters for the data objects. Also the intuitionistic parameter has been retained for the calculation of membership values of element x in a given cluster. UCI repository datasets were used for assessing efficacy of algorithms. The qualified analysis and results depict much consistent performance, where a significant improvement is achieved as compared to intuitionistic fuzzy k-Mode and simulated annealing based intuitionistic fuzzy k-mode. Genetic Algorithm based intuitionistic fuzzy k-Mode is very efficient when clustering is applied on large datasets that are categorical in nature, which proves to be very critical for data mining processes.*

Keywords: *Categorical data, clustering, Data Mining, intuitionistic fuzzy k-Mode, simulated annealing, Genetic Algorithm.*

1. Introduction

Process of deriving incisive deeper knowledge from raw data which results in better decision making is known as data mining. Data Mining lies at the intersection of various fields like machine learning, signal processing, graphics, artificial intelligence, etc. Many methods are used to extract information and recognisable patterns from unstructured data such as clustering, classification, regression, association rules, etc. In this paper clustering algorithm will be worked upon for mining purposes.

The process of grouping a number of entities such that the entities in a particular set are more analogous to each other than to entities in other sets is known as clustering [1]. Most of the raw data available nowadays is without any class values which can be used to classify records properly. Also the class values available do not

have a definite relationship with each other. So in these cases the concept of clustering comes in handy. These techniques tend to increase the intra cluster parity and minimize the inter cluster parity.

Categorical data is the statistical data type consisting of variables that can have only a fixed number of values. Thus each individual is assigned to a particular group or "category". The attributes of various data types are contained by entities in the database. There may be numeric or non-numeric value types. The categorical datasets are therefore classified as shown in Fig. 1.

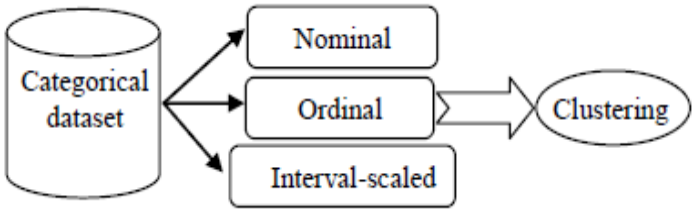

Fig. 1. Clustering categorical data

Clustering can be exercised for any type of data. But clustering numerical data is easier than that of categorical data. There can't be any direct application of the distance metric to the categorical data. So k-Means [2] algorithm which is the most used clustering method is rendered ineffective when applied on categorical data. This is because it fully depends on the distance metric and it can only minimize a cost function which is numerical. So for categorical data, methods which deal with finding the modes are used.

The k-Modes [3] uses simple matching dissimilarity measure and thus is different from the k-Means approach where Euclidean distance method is used. Cluster centers are represented as modes and these modes undergo a change in each iteration of the process, where the most frequent categorical values are put in place of the previous values. A local minima result is guaranteed by these modifications.

Fuzzy k-Modes [4] is an addendum of k-Modes. From categorical data a fuzzy partition matrix is generated within the framework of the fuzzy k-Means algorithm [5, 6]. Its primary concern is to bring out a method to get the fuzzy cluster modes [7, 8] from the categorical entities when the simple matching dissimilarity is applied to them. Confidence to the entities is assigned which makes the fuzzy version better than the k-Modes method. Also, it is quite known that the notion of intuitionistic fuzzy set [9-12] was developed by Atanassov, in which it was indicated that there exists an intuitionistic degree ($\pi_A(x)$) that happens due to lack of knowledge when the membership degree is defined. Using this, an addendum of fuzzy k-Modes known as intuitionistic fuzzy k-Modes was presented.

In this paper a new model has been devised, which is derived on the concepts stated above, called genetic algorithm based intuitionistic fuzzy k-Modes. In this notion the properties of population generation and chromosomes on intuitionistic fuzzy k-Mode has been used. This has been done so as to get better outcome than intuitionistic fuzzy k-Mode while clustering [13, 14] categorical data.

2. Datasets used

UCI machine dataset repository was the primary source of the datasets used in this paper. The datasets used are glass dataset, iris dataset and wine dataset. The description for these datasets is given in Table 1.

Table 1. Datasets description

Data set	Glass dataset	Wine dataset	Iris dataset
Characteristics	Multivariate	Multivariate	Multivariate
Attribute type	Real, Categorical	Real, Integer, Categorical	Real, Categorical
Associated tasks	Classification	Classification	Classification
Number of instances	214	178	150
Number of attributes	9	13	4
Missing values	No	No	No
Class values	1-6	1-3	Iris Setosa, Iris Versicolour, Iris Virginia

3. Notation

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

3.1. Categorical data

It is assumed that the objects to be clustered are stored in a database T defined by a set of attributes $A_1, A_2, \ldots, A_m$. DOM(A_j) is the domain of values described by each attribute A_j and this has an association with a defined semantic data type. Here only two data types are considered which are numeric and categorical, and all other data types in the database are assumed to be linked to these two. Real numbers compose a numeric domain. A domain DOM(A_j) is defined as categorical if it is finite and unordered.

X is represented as a vector $[x_1, x_2, x_3, ..., x_m]$ without ambiguity; m attribute values are contained by every object. Missing values of attribute A_j are denoted by null. Let a group of n objects be denoted by $X = \{X_1, X_2, \ldots, X_n\}$, and $[x_{i1}, x_{i2}, ..., x_{im}]$ represents object X_i. If $x_{i,j} = x_{k,j}$ then $X_i = X_k$; $X_i = X_k$ means, for the attributes $A_1, A_2, \ldots, A_m$, two objects have equal values.

3.2. Intuitionistic fuzzy set

Simultaneous consideration of membership values m and non-membership values n of elements of a group put the notion of intuitionistic fuzzy sets [9, 10] under consideration $\{(x, m_A(x), n_A(x)) \mid x \in X\}$ is an IFS A in X where $m_A : X \to [0,1]$ and $n_A : X \to [0,1]$ such that $0 \le m_A(x) + n_A(x) \le 1$ $\forall x \in X$. An element x has $m_A(x)$ and $n_A(x)$ as membership and non-membership values to set A in X. When $n_A(x) = 1 - m_A(x)$ for every x in set A, then it becomes

a fuzzy set [15-18]. For all Intuitionistic Fuzzy Sets (IFSs) an intuitionistic degree $\pi_A(x)$ was indicated by Atanassov. For every element x in A this arises as a consequence of lack of knowledge and it is given as

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

$m_A(x)$ lie in an interval range $\left[m_A(x) - \pi_A(x),\ m_A(x) + \pi_A(x)\right]$ because of the addition of this hesitation degree.

Intuitionistic Fuzzy Generator (IFG) is used for the construction of IFS. Sugeno's IFG is used in this paper. Intuitionistic fuzzy complement of Sugeno is written as

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

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

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

4. Methods and algorithms

4.1. Distance function

In k-Modes the distance between X and Y where m is the number of attributes that are categorical is given as

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

where $\delta(x_j, y_j) = \begin{cases} 0 & \text{if} \quad x_j = y_j, \\ 1 & \text{if} \quad x_j \neq y_j, \end{cases}$

and X and Y have x_j and y_j as the values of attribute j. This equation is known as simple matching dissimilarity measure. The other name for it is *Hemming distance*. The entities are more dissimilar to each other when there are larger number of mismatches of categorical values between X and Y.

4.2. K-Modes (KM) algorithm

The k-Means clustering algorithm cannot cluster categorical data because of the dissimilarity measure it uses. The k-Modes clustering algorithm is based on k-Means paradigm but removes the numeric data limitation. The k-Modes approach modifies the standard k-Means process for clustering categorical data by replacing 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 the iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minimal

result. Since the k-Means clustering process is essentially not changed, the efficiency of the clustering process is maintained.

k-Modes clustering [3] recurrently find *U* and *Z*. It is an optimization process where a data set *D* is partitioned into *k* clusters. In the process the cost function is minimized:

$$F(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{li}d(Z_l, X_i), \tag{5}$$

subject to

$$\mu_{li}\in\{0,1\},\ \ 1\le l\le k,\ 1\le i\le n, \tag{6}$$

$$\sum_{l=1}^{k}\mu_{li}=1,\ 1\le i\le n, \tag{7}$$

and

$$0<\sum_{i=1}^{n}\mu_{li}<n,\quad 1\le l\le k, \tag{8}$$

where $X_i=\{x_{i,1}, x_{i,2},\ldots, x_{i,m}\}$ denotes *m* categorical attributes and a group of *n* categorical entities are represented as $D=\{X_1, X_2,\ldots, X_n\}$. The current cluster membership of an object is represented by $U=\left[\mu_{li}\right]$ which is a {0, 1} matrix $Z=[Z_1, Z_2,\ldots, Z_k]$ representing 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 (4).

The following three steps are taken by the KM clustering process to cluster a categorical data set ***X*** into *k* clusters.

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 center which has the shortest distance to the object takes the object in its cluster. This step goes on 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 go back to Step 2; otherwise, stop.

KM objective function is minimized by this clustering process:

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

where $U=[\mu_{li}]$ 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 $d(.,.)$.

4.3. Fuzzy KM (FKM) algorithm

The fuzzy k-Modes algorithm was proposed by H u a n g and N g [4] for clustering categorical objects. This algorithm is an extension [19] to k-Modes. Instead of assigning each object to one cluster, the fuzzy k-Modes clustering algorithm calculates a cluster membership degree value for each object to each cluster. Similar to the fuzzy k-Means [5, 6, 20], this is achieved by introducing the fuzziness factor in the objective function. This algorithm has found applications in bioinformatics. It

can improve the clustering result whenever the inherent clusters overlap in a data set. In the fuzzy k-Modes algorithm [21], data D is grouped into k clusters by minimizing the cost function given in (5). The equation for fuzzy membership is

$$\mu_{li} = \begin{cases} 1 & \text{if } X_i = Z_l, \\ 0 & \text{if } X_i = Z_h, h \neq l, \\ 1 / \sum_{j=1}^{k} \left[\frac{d(Z_l, X_i)}{d(Z_h, X_i)} \right]^{\frac{1}{(\alpha-1)}} & \text{if } X_i \neq Z_l \text{ and } X \neq Z_h,\ 1 \leq h \leq k, \end{cases} \tag{9}$$

where α is the weighting component.

Equation (9) gives the fuzzy membership of different objects in a given cluster. The first case is when the object X_i and the cluster center Z_l values are equal and then the membership value of that object in that cluster l is 1. The second case is when the object and another cluster center, some Z_h, are equal and then the membership value of that object in that cluster l is 0. The third case is when the object does not have value equal to any of the cluster center and then the membership value is between 0 and 1. So this equation exemplifies fuzzification.

4.4. Intuitionistic Fuzzy K-Mode (IFKM) algorithm

The Intuitionistic fuzzy k-Modes [21, 22] follows from intuitionistic fuzzy set [23, 24]. While fuzzy sets depend upon graded membership values, intuitionistic fuzzy sets depend upon membership and non-membership values leading to hesitation values associated with every element in the domain. The intuitionistic fuzzy set is explained in Section 3.2. Hesitation value is a new parameter denoted by π. Hesitation value arises due to lack of knowledge in defining membership degree. In this algorithm intuitionistic degree is added to fuzzy k-Mode concept. 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. It gives better result as compared to FKM Algorithm [21]. The steps of the method are as follows:

Step 1. Assign initial cluster centers or modes for c clusters.

Step 2. Between data objects X_i and centroids Z_l the distance d is calculated.

Step 3. The fuzzy partition matrix or membership matrix U is generated as shown by (9).

Step 4. Compute the hesitation matrix π using

$$\pi_{li} = 1 - \mu_{li} - \frac{1-\mu_{li}}{1+\lambda\mu_{li}}. \tag{10}$$

Step 5. Compute the modified membership matrix U' using

$$\mu'_{li} = \mu_{li} + \pi_{li}. \tag{11}$$

Step 6. The X_i with higher relative frequency of categorical attributes is chosen to be the new representative, i.e., center or mode.

Step 7. By using Steps 2-5 the new partition matrix is calculated.

Step 8. If $\left\| U'^{(r)} - U'^{(r+1)} \right\| < \varepsilon$ then stop. Else repeat from Step 4.

The objective function of IFKM [22] 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

$$J_{\text{IFKM}} = \sum_{l=1}^{k}\sum_{i=1}^{n} \mu_{li}^{*m} d_{li}^2 + \sum_{l=1}^{k} \pi_l^* e^{1-\pi_l^*}, \quad \text{where } \pi_l^* = \frac{1}{N}\sum_{i=1}^{n} \pi_{li}, \quad i \in [1, N], \tag{12}$$

and N is the number of objects.

4.5. Genetic Algorithm based Intuitionistic Fuzzy K-Mode (GAIFKM)

John Holland first used the term Genetic Algorithm (GA) (see [25, 26]). The concept of GA is based on the philosophy of natural selection or "survival of the fittest" and genetics which are inspired by biological structures and their evolution. The use of GA is very much effective while the search space is large, complex and multimodal. Unlike traditional search techniques, it works with the coding of problem variables instead of variable themselves. Moreover, it can search simultaneously from multiple points and this fact makes GA parallel in nature. This also helps increasing the probability of avoiding the issue of getting trapped into local optimal solution. According to the GA, the decision variables of the search problem are encoded into a finite length string of symbols of certain cardinality. The symbols are called genes and the values of genes are called alleles, whereas the string is referred as chromosome. GA [25] starts with the random population of chromosomes. Each chromosome is evaluated based on some defined fitness value and then takes part into selection process where the chromosomes having better fitness are given more chance to reproduce than the others. Thereafter, crossover and mutation operators are applied on the selected chromosomes. This preserves the important information and helps to achieve the good solution for next generation. This process continues until some termination condition is reached.

4.5.1. Selection process

During selection process [27], chromosomes from the parent populations are selected based on the fitness values and an intermediate population, called mating pool is maintained. These selected chromosomes will further take part into subsequent processes like crossover and mutation. Three selection methods, namely roulette wheel selection, stochastic universal sampling and binary tournament selection are very much popular and mostly used. In this paper binary tournament selection method has been used for GA. Further explanation for this is given in Section 4.5.5.

4.5.2. Crossover process

In this process two parent chromosomes produce two new chromosomes called offspring. Offspring chromosomes might get the best characteristics from both the parents and become better than parents. Crossover does not occur all the time. It occurs based on some user defined probability. There are single point crossover, two-point and uniform crossover. Single point crossover has been used in this paper. According to it, crossover process selects a crossover point over a chromosome

randomly. Thereafter, it interchanges the two parent chromosomes at this point to produce offspring chromosomes.

4.5.3. Mutation process

Mutation is another genetic operator [28, 29]. It alters one or more genes in a chromosome and can produce entirely a new chromosome than its initial state. This new chromosome is added to the population with the assumption that the new chromosome might give the better solution. This process also follows user defined probability for occurrence. The probability of mutation is normally kept less (as low as 0.01) than crossover.

4.5.4. Elitism

During the GA process, there is a chance of losing the best solution in next generation unless the best solution is stored so far in a safe place. The process which ensures it is called *elitism* [30, 31]. Sometimes, the worst solution so far is replaced by the best solution so far in the population.

4.5.5. Main algorithm

So here intuitionistic fuzzy k-Mode has been extended by adding the concept of genetic algorithm to it. The procedure is as follows:

Input: X, the dataset
k, the number of cluster
maxGen, maximum Generation
PS, Population Size
P_{cr}, crossover Probability
P_{mu}, mutation Probability
Output: $[\mu_{li}]$ where $1 \leq l \leq k$ and $1 \leq i \leq n$.

Step 1. Select random k objects from dataset for k cluster mode to encode as chromosome. Z_l is the cluster mode for $l = 1, 2,..., k$.

Step 2. Generate initial population of size PS.

Step 3. **repeat**

Step 4. Calculate μ_{li} for all n objects using (9).

Step 5. Classify objects using algorithm given in Section 4.4 and update $[\mu_{li}]$.

Step 6. Calculate fitness value using (12) for each chromosome in population.

Step 7. Update each chromosome in population with new mode using (11).

Step 8. Selection using *tournament selection* strategy.

Step 9. Perform crossover with probability P_{cr}.

Step 10. Perform mutation with probability P_{mu}.

Step 11. **until** maxGen is reached

In GAIFKM, the chromosome is encoded similarly like a string. Each chromosome indicates a probable solution. The fitness of a chromosome indicates the degree of goodness of the solution it represents. In this paper, the J_{IFKM} is used as the objective function as defined in (12). The objective is therefore to minimize the J_{IFKM} for achieving optimal clustering. Given a chromosome, the modes encoded in it are

first extracted and then the Equations (9), (12) and (11) are used to compute fuzzy membership matrix, J_{IFKM} and updating modes respectively.

Thereafter, the popular genetic operations *selection*, *crossover* and *mutation* are used. Here the tournament selection strategy is adopted. In tourrnament selection strategy some individuals are selected randomly and each selected individual competes against each other. The individual with best fitness wins and is included in the next generation population. The tournament size which is the number of individuals competing each other in each tournament is normally set to 2. The tournament selection strategy gives all individuals a chance to be selected and thus preserves diversity. After selection, the selected chromosomes are used for crossover operation. Conventional single point crossover with probability P_{cr} has been performed for generating the new offspring. Subsequently, mutation, with probability P_{mu} was carried out. A chromosome is selected for mutation. The gene position that will undergo mutation is chosen randomly. Subsequently, the categorical value of that position is replaced by another one chosen randomly from the corresponding categorical domain. Elitism strategy [30, 31] has been implemented by preserving the best chromosome in a separate location outside the population. At the end this provides the best chromosome consisting of modes of final clusters. All these processes are repeated for a maximum number of generation maxGen.

5. Criteria to be used for evaluation

One of the most basic performance analysis indexes are Davis-Bouldin (DB) and Dunn (*D*) indexes [32, 33] . They help in evaluating the efficiency of clustering. The number of clusters required determines the results.

5.1. Davis-Bouldin index

The ratio of sum of intra-cluster distance to inter-cluster distance is known as DB index [34]. It is formulated as

$$\text{DB} = \frac{1}{k}\sum_{l=1}^{k}\sum_{o\neq l, o=1}^{k} \max\left\{\frac{S(Z_l)+S(Z_o)}{d(Z_l, Z_o)}\right\} \text{ for } 1< l, o < k, \tag{13}$$

S is the sum of intra-cluster distance. So *S* = Sum (Distance between any two objects taken at a time which are present in the same cluster); *d* is the distance between the two clusters. The objective of this index is to reduce the within cluster distance and increase the between cluster separation. Therefore a low DB index indicates that the clustering procedure is good.

5.2. Dunn (*D*) index

The *D* index [35] is similar to DB index. Compact and separated clusters are identified by it. Computation is done by using

$$\text{Dunn} = \min_l\left\{\min_{o\neq l}\left\{\frac{d(Z_l, Z_o)}{\max_b S(Z_b)}\right\}\right\} \text{ for } 1 < o, l, b < k. \tag{14}$$

Its aim is to maximize the inter-cluster distance and reduce the intra-cluster distance is. Hence a procedure is more efficient if the value of the D index is greater.

5.3. Clustering accuracy

The accuracy of a clustering process is defined as

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

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

6. Results and analysis

To assess the efficacy and efficiency of the genetic algorithm based intuitionistic fuzzy k-Modes method and compare it with the intuitionistic fuzzy k-Modes algorithm, several tests of these algorithms were carried out.

The datasets used were the glass dataset, iris dataset and wine dataset. All the three datasets directly have been taken from UCI repository. No changes like removing some redundant rows, cleaning the data or removing some attributes, have been made to the datasets.

For the dataset the two clustering algorithms to cluster it have been used. For IFKM algorithm $\lambda = 2$. The record X_i was assigned to the L-th cluster if $\mu_{li} = \max_{1 \le h \le k} \{\mu_{hi}\}$. For genetic algorithm based intuitionistic fuzzy k-Mode the different values taken for the constants used in the algorithm were $P_{\mathrm{cr}} = 0.8$, $P_{\mathrm{mu}} = 0.1$ and maxGen = 100.

The number of clusters that has been taken are 6, 3, and 3 for glass, iris and wine dataset respectively. This is because these datasets contain these many different values in the decision attribute. Final modes of these clusters produced on applying the two algorithms are given in the table below. These modes are non-identical. This suggests that the genetic algorithm based intuitionistic fuzzy k-Modes and intuitionistic fuzzy k-Modes algorithms indeed produce different clusters.

6.1. Modes of the clusters

In this section the cluster centers have been computed for intuitionistic fuzzy k-Mode and genetic algorithm based intuitionistic fuzzy k-Modes for three data sets; glass dataset, iris dataset and wine dataset to show the superiority of genetic algorithm based intuitionistic fuzzy k-Mode over the intuitionistic Fuzzy k-Mode Algorithm. Z_l represents the different clusters or cluster modes. It is denoted in the first column from Tables 2-6. The first row in Tables 2-6 contains the different attributes present in the dataset. So for every attribute and decision class a cluster center is chosen according to the applied algorithm.

6.1.1. Glass dataset

Table 2 shows the cluster centers or modes obtained by using the Intuitionistic Fuzzy K-Mode Algorithm (IFKMA).

Table 2. Modes for glass dataset on applying IFKM

Z_l	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 3 shows the cluster modes obtained by using the GA based IIFKM.

Table 3. Modes for glass dataset on applying genetic based IIFKM

Z_l	1	2	3	4	5	6	7	8	9
1	1.5165	12.98	2.14	1.70	72.92	0.07	9.3	0.12	0.38
2	1.5165	14.15	2.14	1.70	71.31	0.41	9.3	3.13	0.04
3	1.5165	14.15	2.14	1.70	74.42	0.2	9.3	0.12	0.38
4	1.5165	14.15	2.14	1.70	72.91	2.71	9.3	0.12	0.02
5	1.5165	12.98	2.96	2.10	72.91	0.93	13.3	0.11	0.25
6	1.5166	12.98	3.65	0.62	72.93	0.54	8.03	0.11	0.04

6.1.2. Iris dataset

Table 4 shows the cluster modes obtained by using the two algorithms.

Table 4. Modes for iris dataset on applying the two algorithms

Z_l	GA based on IIFKM				IIFKM			
	1	2	3	4	1	2	3	4
1	6.8	4.3	4.2	0.5	4.3	2	1.1	0.6
2	4.3	2.1	1.1	0.5	7	4.1	1	0.6
3	7.2	2.3	6.3	0.6	7	2	3.6	0.5

6.1.3. Wine dataset

Tables 5 and 6 show the cluster modes obtained by using the IFKMA and GA based IFKM, respectively.

Table 5. Modes for wine dataset on applying IFKM Algorithm

Z_l	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 6. Modes for wine dataset on applying GAIFKM

Z_l	1	2	3	4	5	6	7	8	9	10	11	12	13
1	12.81	3.85	2.3	20	105	1.2	0.77	0.34	1.03	2.2	0.64	3.1	407
2	11.04	3.87	2	18	82	1.64	0.77	0.14	0.85	8.2	0.64	2.48	851
3	13.30	2.82	2.8	25	82	1.88	1.91	0.42	2	10.25	1.42	2.82	409

So Tables 2 up to 6 give the final cluster modes on application of the respective algorithms on the datasets. We see that cluster modes are different when we apply IFKM and GAIFKM on the same dataset. This is because in GAIFKM we have used various concepts related to genetics like tournament selection strategy, crossover, mutation and elitism which gives better chances of getting good cluster modes. The cluster modes found out by using GAIFKM are much better in the form that clusters made by them are more resonant. This is because they increase the inter-cluster distance and reduce the intra-cluster distance as indicated by the results of DB and D-index values in section below. Also the Tables 2 to 6 can be used for verification purpose by someone who wants to further work on the algorithm and extend it.

6.2. DB and *D*-index values

Now the DB and *D*-index of the two algorithms is calculated according to the formulas given in Section 5. The representation for this has been made with the help of a table shown below which clearly indicate that genetic algorithm based intuitionistic fuzzy k-Mode is better than intuitionistic fuzzy k-Mode.

Table 7. DB and D-index values

Datasets	IFKM		GAIFKM	
	DB	*D*	DB	*D*
Glass	11.1	0.1111	4.13	0.39
Iris	2.667	0.75	1.698	1.05
Wine	2.1667	0.9231	2.146	0.95

6.3. Accuracy

Now the clustering accuracy of the two algorithms has been calculated. The accuracies are as follows:

Table 8. Accuracy of clustering

Datasets	IFKM	GAIFKM
Glass	0.67	0.74
Iris	0.555	0.595
Wine	0.624	0.68

The accuracy of application of genetic algorithm based intuitionistic fuzzy k-Mode on the three datasets is much more than that of intuitionistic fuzzy k-Mode. So the results obtained in Table 8 clearly justify that genetic algorithm based intuitionistic fuzzy k-Mode is a much better method for categorical data based clustering than intuitionistic fuzzy k-Mode.

7. Conclusion

When it comes to real-world databases, categorical datasets have become a necessity. However, for clustering massive categorical data only a few efficient algorithms are available. The introduction of the k-modes type algorithm and its extension in the form of fuzzy k-Modes algorithm for clustering categorical objects, was the required impetus 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, genetic algorithm based intuitionistic fuzzy k-Mode technique has been introduced. This process used other parameters which were very much different from those used in intuitionistic fuzzy k-Mode. The additional computation is required in the elimination process which is iterative. But it does not affect the complexity which remains linear. The application of the algorithm on the three datasets have shown that a large number of initial modes are used which enhance the performance of the method. This happens 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 it is perceived that the genetic algorithm based intuitionistic fuzzy k-Mode method performs better than the intuitionistic fuzzy k-Mode algorithm as demonstrated in this paper. The findings possess a major application for data mining where the uncertain objects on the boundary are at times more interesting than entities which can be grouped by using clustering techniques with certainty.

8. Scope for future work

Better clusters can be formed by using a much better distance function. The cluster formed in the end depends heavily on initial cluster taken. Thus finding a way to choose better initial cluster can lead to better cluster formation. Also different threshold value provides different set of cluster. So according to our application it can be changed for better result. In addition to this, the crossover probability, mutation factor, and number of iterations could be varied and a concrete relation between these things could help get much better clusters.

References

1. M a c Q u e e n, J. Some Methods for Classification and Analysis of Multivariate Observations. – In: Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. **1**, 1967, pp. 281-297.
2. H a r t i g a n, J. A., M. A. W o n g. Algorithm AS 136: A k-Means Clustering Algorithm. – Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. **28**, 1979, No 1, pp. 100-108.
3. H u a n g, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. – Data Mining and Knowledge Discovery, Vol. **2**, 1998, No 3, pp. 283-304.
4. H u a n g, Z., M. K. N g. A Fuzzy k-Modes Algorithm for Clustering Categorical Data. – IEEE Transactions on Fuzzy Systems, Vol. **7**, 1999, No 4, pp. 446-452.

5. R u s p i n i, E. H. A New Approach to Clustering. – Information and Control, Vol. **15**, 1969, No 1, pp. 22-32.
6. R u s p i n i, E. H. Numerical Methods for Fuzzy Clustering. – Information Sciences, Vol. **2**, 1970, No 3, pp. 319-350.
7. Y a n g, M.-S. A Survey of Fuzzy Clustering. – Mathematical and Computer Modelling, Vol. **18**, 1993, No 11, pp. 1-16.
8. P e l e k i s, N., D. K. I a k o v i d i s, E. E. K o t s i f a k o s, I. K o p a n a k i s. Fuzzy Clustering of Intuitionistic Fuzzy Data. – International Journal of Business Intelligence and Data Mining, Vol. **3**, 2008, No 1, pp. 45-65.
9. A t a n a s s o v, K. T. More on Intuitionistic Fuzzy Sets. – Fuzzy Sets and Systems, Vol. **33**, 1989, No 1, pp. 37-45.
10. A t a n a s s o v, K. T. Intuitionistic Fuzzy Sets. – Fuzzy Sets and Systems, Vol. **20**, 1986, No 1, pp. 87-96.
11. Z h a n g, H., Z. X u, Q. C h e n. On Clustering Approach to Intuitionistic Fuzzy Sets. – Control and Decision, Vol. **22**, 2007, No 8, p. 882.
12. X u, Z., J. C h e n, J. W u. Clustering Algorithm for Intuitionistic Fuzzy Sets. – Information Sciences, Vol. **178**, 2008, No 19, pp. 3775-3790.
13. C h a i r a, T. A Novel Intuitionistic Fuzzy C Means Clustering Algorithm and Its Application to Medical Images. – Applied Soft Computing, Vol. **11**, 2011, No 2, pp. 1711-1717.
14. C h a i r a, T., A. P a n w a r. An Atanassov's Intuitionistic Fuzzy Kernel Clustering for Medical Image Segmentation. – International Journal of Computational Intelligence Systems, Vol. **7**, 2014, No 2, pp. 360-370.
15. D u b o i s, D. J. Fuzzy Sets and Systems: Theory and Applications. – Academic Press, Vol. **144**, 1980.
16. K l i r, G., B. Y u a n. Fuzzy Sets and Fuzzy Logic. Vol. **4**. New Jersey, Prentice Hall, 1995.
17. Z a d e h, L. A. Fuzzy Sets. – Information and Control, Vol. **8**, 1965, No 3, pp. 338-353.
18. K l i r, G. J., T. A. F o l g e r. Fuzzy Sets, Uncertainty, and Information. 1988.
19. D e s c h r i j v e r, G., E. E. K e r r e. On the Relationship between Some Extensions of Fuzzy Set Theory. – Fuzzy Sets and Systems, Vol. **133**, 2003, No 2, pp. 227-235.
20. B e z d e k, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. – Springer Science & Business Media, 2013.
21. T r i p a t h y, B. K., A. G o y a l, A. S. P a t r a. Clustering Categorical Data Using Intuitionistic Fuzzy k-Mode. – International Journal of Pharmacy and Technology, Vol. **8**, 2016, No 3, pp. 16688-16701.
22. T r i p a t h y, B. K., A. G o y a l, A. S. P a t r a. A Comparative Analysis of Rough Intuitionistic Fuzzy k-Mode for Clustering Categorical Data. – Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol. **7**, 2016, No 5, pp. 2787-2802.
23. X i e, N., Z. L i, G. Z h a n g. An Intuitionistic Fuzzy Soft Set Method for Stochastic Decision-Making Applying Prospect Theory and Grey Relational Analysis. – Journal of Intelligent & Fuzzy Systems, Vol. **33**, 2017, No 1, pp. 15-25.
24. D a s, S., D. G u h a. Similarity Measure of Intuitionistic Fuzzy Numbers and Its Application to Clustering. – International Journal of Mathematics in Operational Research, Vol. **10**, 2017, No 4, pp. 399-430.
25. G o l d b e r g, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addion Wesley, 1989, p. 102.
26. S a h a, I., A. M u k h o p a d h y a y. Genetic Algorithm and Simulated Annealing Based Approaches to Categorical Data Clustering. – In: Proc. of IEEE Region 10 and the 3rd International Conference on Industrial and Information Systems, 2008, pp. 1-6.
27. C h e n g, C. H., W. K. L e e, K. F. W o n g. A Genetic Algorithm-Based Clustering Approach for Database Partitioning. – IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. **32**, 2008, No 3, pp. 215-230.
28. C o w g i l l, M. C., R. J. H a r v e y, L. T. W a t s o n. A Genetic Algorithm Approach to Cluster Analysis. – Computers & Mathematics with Applications, Vol. **37**, 1999, No 7, pp. 99-108.
29. S h e i k h, R. H., M. M. R a g h u w a n s h i, A. N. J a i s w a l. Genetic Algorithm Based Clustering: A Survey. – In: Proc. of First International Conference on Emerging Trends in Engineering and Technology, 2008, pp. 314-319.

30. D e b, K., S. A g r a w a l, A. P r a t a p, T. M e y a r i v a n. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. – In: Proc. of International Conference on Parallel Problem Solving from Nature, 2000, pp. 849-858.
31. D e b, K., A. P r a t a p, S. A g a r w a l, T. M e y a r i v a n. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. – IEEE Transactions on Evolutionary Computation, Vol. **6**, 2002, No 2, pp. 182-197.
32. M a u l i k, U., S. B a n d y o p a d h y a y. Performance Evaluation of Some Clustering Algorithms and Validity Indices. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **24**, 2002, No 12, pp. 1650-1654.
33. X i e, X. L., G. B e n i. A Validity Measure for Fuzzy Clustering. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **13**, 1991, No 8, pp. 841-847.
34. D a v i e s, D. L., D. W. B o u l d i n. A Cluster Separation Measure. – IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, No 2, pp. 224-227.
35. B e z d e k, J. C., N. R. P a l. Cluster Validation with Generalized Dunn's Indices. – In: Proc. of Artificial Neural Networks and Expert Systems, 2nd New Zealand International Two-Stream Conference, 1995, pp. 190-193.