IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 1, July 2013

# A Cluster Centres Initialization Method for Clustering Categorical Data Using Genetic Algorithm

Kusha Bhatt
kusha.bhatt@gmail.com

Prof. Pankaj Dalal
pkjdalal@gmail.com

Prof Avinash Panwar
avinashpanwar@gmail.com

*Abstract:* **The leading partitioned clustering technique, k-modes, is among the most computationally efficient clustering methods for categorical data. However, the k-modes clustering algorithm performance, which converges to numerous local minima strongly depends on initial cluster centres. Currently, most methods of initialization cluster centres are mainly for numerical data. Due to lack of geometry for the categorical data, these methods used in cluster centres initialization for numerical data are not applicable to categorical data. This research proposes a novel initialization method for categorical data which is implemented to the k-modes algorithm using genetic algorithm. The method integrates the distance and the density together to select initial cluster centres and overcomes shortcomings of the existing initialization methods for categorical data. Genetic algorithm is used here to optimize cluster centre initialization in traditional K-mode algorithm in order to find best results.**

*Keywords:* **Clustering, k-modes, Genetic Algorithm**

## I. Introduction

We often need to partition a set of objects in databases into homogeneous groups or clusters. This is a fundamental operation in data mining. It is a key feature to be executed in a number of tasks, such as classification (unsupervised) aggregation and segmentation or dissection. The problem of clustering is defined as the partitioning the data set consisting of n points embedded in m-dimensional space into k distinct set of clusters, such that the data points within the same cluster are more similar to each other than to data points in other clusters. The three sub-problems that are mainly resolved by the clustering process are (i) defining a similarity measure to judge the similarity (or distance) between different elements (ii) implementing an efficient algorithm to discover the clusters of most similar elements in an unsupervised way and (iii) derive a description that can characterize the elements of a cluster in a succinct manner.

Clustering based on k-means is closely related to a number of other clustering and location problems. The approach in K-means method use the Euclidean k-medians (or the multisource Weber problem). In this method the sum of distances should be minimized to the center and the geometric-center problem in which it is required to minimize the maximum distance from every point to its closest center. We cannot find any unique solution for both the problems and also some formulations are NP-hard. An asymptotically efficient approximation for the k-means clustering problem has been presented by Matousek, but the large constant factors suggest that it is not a good candidate for practical implementation.

One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the k-means algorithm.

K-means does not guarantee unique clustering because we get different results with randomly chosen initial cluster centers and hence the results cannot be relied with confidence. The K-means algorithm gives better results only when the initial partitions are close to the final solution. Several attempts have been reported to generate K-prototype points that can be used as initial cluster centers. A recursive method for initializing the means by running K clustering problems is discussed by Dudaand Hart. Bradley et al reported that the values of initial means along any one of the m coordinate axes are determined by selecting the K densest "bins" along that coordinate. Bradley and Fayyad [1998] proposed a procedure that refines the initial point to a point likely to be close to the modes of the joint probability density of the data. Mitra-et-al suggested a method to extract prototype points based on Density Based Multi-scale Data Condensation. Khan and Ahmad presented an algorithm to compute initial cluster centers for K-means clustering algorithm. Their algorithm is based on two experimental observations that some of the patterns are very

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center. The initial cluster centers computed by using their methodology are found to be very close to the desired cluster centers with improved and consistent clustering results [1].

K-Modes is an extension of K-Means clustering algorithm, but the working principle of both is same. Instead of means, the concept of modes are used. By varying the dissimilarity measure, fuzzy K-Modes, K-representative and K-histogram are developed. In fuzzy K-Modes, instead of hard centroid, soft centroid is used [2]. In K-representative algorithm, the measure relative frequency is used. Frequency of attribute value in the cluster divided by cluster length is used as a measure in K-representative. In K-histogram, proportion of attribute value in the data set is considered. K-Modes is extended with fuzzy, genetic and fuzzy-genetic concepts. As the proposed measure is similar to K-Modes and K-representative, we compared the proposed measure with K-Modes and K-representative.

## II. Related Work

Various clustering algorithms have been reported to cluster categorical data. He, Z. et al(2005) [3] proposed a cluster ensemble for clustering categorical data. Ralambondrainy (1995) [4] presented an approach by using k-means algorithm to cluster categorical data. The approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and treat the binary attributes as numeric in the k-means algorithm. Gowda and Diday (1991) [5] used other dissimilarity measures based on “position”, “span” and “content” to process data with categorical attributes. Huang (1998) [6] proposed k-modes clustering which extend the k-means algorithm to cluster categorical data by using a simple matching dissimilarity measure for categorical objects. Recently, Chaturvedi, et al (2001) [7] also presented k-modes which used a nonparametric approach to derive clusters from categorical data using a new clustering procedure. Huang, Z (2003) [8] has demonstrated the equivalence of the two independently developed k-modes algorithm given in two papers which done by Huang, Z. (1998) [6] and Chaturvedi, et al (2001)[7]. Then, San, O.M., et al (2004) [9] proposed an alternative extension of the k-means algorithm for clustering categorical data which called k-representative clustering. Huang in Huang (1998) suggested to select the first k distinct objects from the data set as the initial k modes or assign the most frequent categories equally to the initial k modes. Though the methods are to make the initial modes diverse, an uniform criteria is not given for selecting k initial modes in Huang (1998). Sun, Zhu, and Chen (2002) introduces an initialization method which is based on the frame of refining. This method presents a study on applying Bradley’s iterative initial-point refinement algorithm (Bradley &Fayyad, 1998) to the k-modes clustering, but its time cost is high and the parameters of this method are plenty which need to be asserted in advance. In Coolcat algorithm (Barbara, Couto, & Li, 2002),the MaxMin distances method is used to find the k most dissimilar data objects from the data set as initial seeds. However, the method only considers the distance between the data objects, by which outliers maybe be selected. Cao, Liang, and Bai (2009) and Wu, Jiang, and Huang (2007) integrated the distance and the density together to propose a cluster centers initialization method, respectively. The difference between the two methods is the definition of the density of an object [10]. Wu used the total distance between an object and all objects from data set as the density of the object. Due to the fact that the time complexity of calculating the densities of all objects is O(n2), it limits the process in a sub-sample data set and uses are fining framework. But this method needs to randomly select sub-sample, so the sole clustering result cannot be guaranteed.

## III. K-Modes 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 whilst preserving its efficiency. The K-modes algorithm extends K-means paradigm to cluster categorical data by removing the limitation imposed by K-means through following modifications:

- Using a simple matching dissimilarity measure or the hamming distance for categorical data objects
- Replacing means of clusters by their modes

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

The simple matching dissimilarity measure [Jain and Dubes, 1988] can be defined as following. Let X and Y be two categorical data objects described by F categorical attributes. The dissimilarity measure d( ) X,Y between X and Y can be defined by the total mismatches of the corresponding attribute categories of two objects. Smaller the number of mismatches, more similar the two objects are. Mathematically, we can say:

$$d(X,Y) = \sum_{j=1}^{F} \partial(x_j, y_j)$$

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

d(X,Y) gives equal importance to each category of an attribute.

Let Z be a set of categorical data objects described by categorical attributes, A1, A2….. $A_F$, a mode of $Z=\{Z_1, Z_2, \ldots, Z_n\}$ is a vector $Q=[q_1, q_2, \ldots, q_F]$ that minimizes

$$D(Z,Q) = \sum_{i=1}^{n} d(Z_i, Q)$$

Here, Q is not necessarily an element of Z. When the above is used as the dissimilarity measure for categorical data objects, the cost function becomes

$$C(Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{F} \partial(z_{ij}, q_{lj})$$

Where $Q_1=[q_{l1}, q_{l2}, \ldots, q_{lm}] \in Q$

The K-modes algorithm minimizes the cost function defined in above equation.

The K-modes algorithm consists of the following steps: -

a. Select K initial modes, one for each of the cluster.
b. Allocate data object to the cluster whose mode is nearest to it according to equation 2.1.
c. Compute new modes of all clusters.
d. Repeat step 2 to 3 until no data object has changed cluster membership.

Like the k-means algorithm the k-modes algorithm also produces locally optimal solutions that are dependent on the initial modes and the order of objects in the data set. In our current implementation of the k-modes algorithm we include two initial mode selection methods. The first method selects the first k distinct records from the data set as the initial k-modes. The second method is implemented with the following steps.

*1. Calculate the frequencies of all categories for all attributes and store them in a category array in descending order of frequency as shown in figure 1. Here, $c_i, j$ denotes category $i$ of attribute $j$ and $f(c_i, j) \geq f(c_{i=i+1}, j)$ where $f(c_i, j)$ is the frequency of category $c_i, j$*

*2. Assign the most frequent categories equally to the initial k modes. For example in figure 1, assume $k = 3$.*

*We assign* $Q_1 = [q_{1,1} = c_{1,1}, q_{1,2} = c_{2,2};\ q_{1,3} = c_{3,3}, q_{1,4} = c_{1,4}]$

$$\begin{Bmatrix} c_{1,1} & & c_{1,3} & \\ c_{2,1} & c_{1,2} & c_{2,3} & c_{1,4} \\ c_{3,1} & c_{2,2} & c_{3,3} & c_{2,4} \\ c_{4,1} & & c_{4,3} & c_{3,4} \\ & & c_{5,3} & \end{Bmatrix}$$

*Figure 1. The category array of a data set with four attributes having 4, 2, 5, 3 categories, respectively.*

$$Q_2 = [q_{2,1} = c_{2,1}, q_{2,2} = c_{1,2}, q_{2,3} = c_{4,3}, q_{2,4} = c_{2,4}]$$

$$Q_3 = [q_{3,1} = c_{3,1}, q_{3,2} = c_{2,2}, q_{3,3} = c_{1,3}, q_{3,4} = c_{3,4}]$$

*3. Start with, $Q_1$. Select the record most similar to $Q_1$ and replace $Q_1$ with the record as the first initial mode. Then select the record most similar to $Q_2$ and replace $Q_2$ with the record as the second initial mode. Continue this process until $Q_k$ is replaced. In these selections $Q_1 \neq Q_t$ for $l \neq t$.*

*Step 3 is taken to avoid the occurrence of empty clusters. The purpose of this selection method is to make the initial modes diverse, which can lead to better clustering results.*

## IV. Genetic Approach

It start with a random configuration of cluster Genetic K mode Algorithm Implementation and Analysis centres. In every iteration, each pattern is assigned to the cluster whose center is the closest center to the pattern among all the cluster centres.

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

The cluster centres in the next iteration are the centroids of the patterns belonging to the corresponding clusters. The algorithm is terminated when there is no reassignment of any pattern from one cluster to another or the variation measure ceases to decrease significantly after an iteration. A major problem with this algorithm is that it is sensitive to the selection of initial partition and may converge to a local minimum of variation if the initial partition is not properly chosen and also with the value of centroid for the respected cluster in order to make partitions optimum. That's why we are going to select the number of partitions and the value of centroids optimally through genetic algorithm. Also, cluster centers initialization method in K-mode Algorithm is optimized through Genetic Algorithm. K-mode Algorithm is used by genetic algorithm as a fitness function and after that genetic algorithm will decide the optimum numbers and values of centroid based on the fitness function and also optimize the cluster centers initialization method.

A typical genetic algorithm requires:

- Genetic representation of the solution domain
- Fitness function to evaluate the solution domain

The basic genetic algorithm is as follows:

- **Start** - Genetic random population of n chromosomes (suitable solutions for the problem)
- **Fitness** - Evaluate the fitness f(x) of each chromosome x in the population
- **New population** - Create a new population by repeating following steps until the New population is complete
- **Selection** - Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to get selected).
- **Crossover** - With a crossover probability, cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
- **Mutation** - With a mutation probability, mutate new offspring at each locus (position in chromosome)
- **Accepting** - Place new offspring in the new population.
- **Replace** - Use new generated population for a further sum of the algorithm.
- **Test** - If the end condition is satisfied, stop, and return the best solution in current population.
- **Loop** - Go to step2 for fitness evaluation.

A basic flow of Genetic Algorithm can be understand by the following figure:

## V. Proposed Work

A categorical domain contains only singletons. Combinational values are not allowed. A special value, denoted by,∈ is defined on all categoricaldomains and used to represent missing values. To simplify the dissimilarity measure we donot consider the conceptual inclusion relationships among values in a categorical domain such that car and vehicle are two categorical values in a domain and conceptually a car is also a vehicle. However, such relationships may existing real world databases.

An object X is logically represented as a conjunction of attribute-value pairs

$$[A_1 = x_1] \wedge [A_2 = x_2] \wedge .. \wedge [A_m = x_m]$$

Where $x_j \in DOM(A_j)$for $1 \leq j \leq m$. An attribute-value pair $[A_j = x_j]$ is called aselector. Without ambiguity we represent X as a vector

$$x_1^r, x_2^r, \ldots . x_p^r, x_1^r, x_{p+1}^c, \ldots . x_m^c$$

Where the first p elements are numeric values and the rest are categorical values. If X has only one type of value, it is simplified as

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

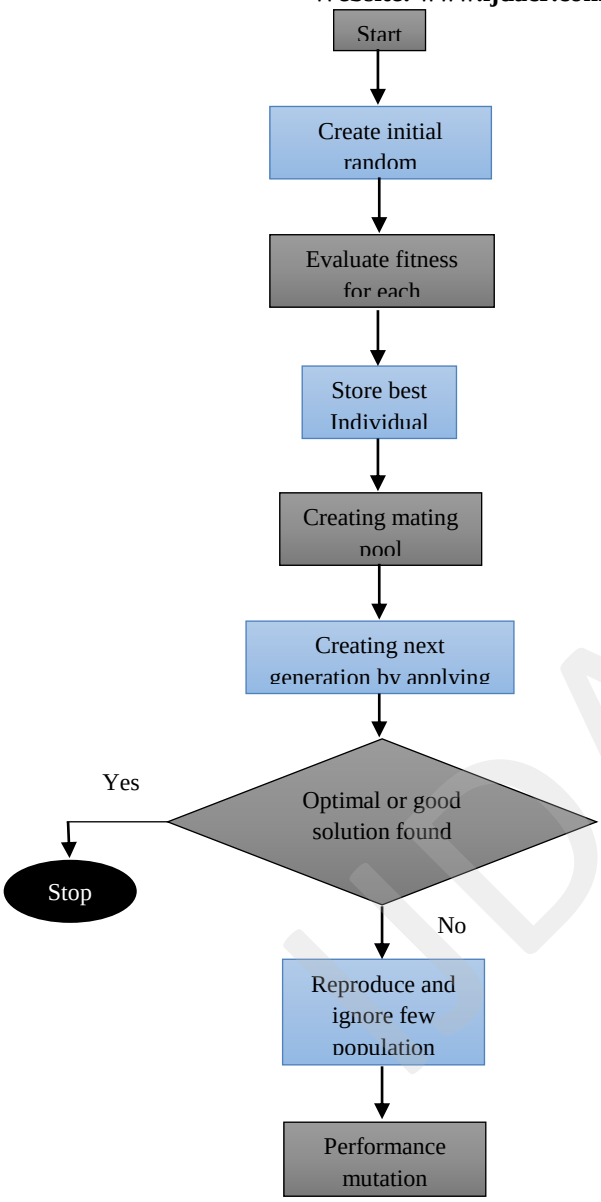


Figure 1: Flow Chart of GA

$[x_1, x_2, \dots x_m]$

X is called a numeric object if it has only numeric values. It is called a categorical object if it has only categorical values. It is called a mixed-type object if it contains both numeric and categorical values.
We consider every object has exactly m attribute values and do not allow numeric attributes to have missing values. If the value of a categorical attribute $A_j^c$ is missing for an object X,then $A_j^c = \in$.

Let $X = \{X_1, X_2, \dots, X_n\}$ be a set of n objects. Object $X_i$ is represented as $[x_{i,1}, x_{i,2}, \dots, x_{i,m}]$.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 notmean that $X_i$ and $X_k$ are the same object in the real world database. It means the two objects have equal values for the attributes $A_1, A_2, \dots, A_n$. For example, two patients in a data setmay have equal values for the attributes Age, Sex, Disease and Treatment. However, theyare distinguished in the hospital database by other attributes such as ID and Address which were not selected for clustering.

## VI. Results and Discussion

In this section, in order to evaluate the performance and scalability of the proposed initialization method, some standard data sets are downloaded from the UCI Machine Learning Repository (2010). All missing attribute values are treated as special values.

**Soybean Large Dataset**

Relevant Information Paragraph:

There are 19 classes, only the first 15 of which have been used in prior work. The folklore seems to be that the last four classes are unjustified by the data since they have so few examples.

**Zoo Dataset**

Relevant Information:

A simple database containing 17 Boolean-valued attributes. The "type" attribute appears to be the class attribute. Here is a breakdown of which animals are in which type: (I find it unusual that there are 2 instances of "frog" and one of "girl"!)

Class# Set of animals:

1 (41) aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruit bat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sea lion, squirrel, vampire, vole, wallaby,wolf2 (20) chicken, crow, dove, duck,

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren 3 (5) pitviper, seasnake, slowworm, tortoise, tuatara 4 (13) bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna 5 (4) frog, frog, newt, toad 6 (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp 7 (10) clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

Number of Instances: 101

Number of Attributes: 18 (animal name, 15 Boolean attributes, 2 numerics)

**Performance Evaluation:**

To evaluate the performance of clustering results, an evaluation method is introduced. If a data set contains k classes for a given clustering, let $a_i$ denote the number of data objects that are correctly assigned to class $C_i$, Let $b_i$ denote the data objects thatare incorrectly assigned to the class $C_i$, and let ci denote the data objects that are incorrectly rejected from the class $C_i$. The accuracy, precision and recall are defined as follow:

$$AC = \frac{\sum_{i=1}^{k} a_i}{|U|}$$

$$PR = \frac{\sum_{i=1}^{k} \left( \frac{a_i}{a_i + b_i} \right)}{k}$$

$$RE = \frac{\sum_{i=1}^{k} \left( \frac{a_i}{a_i + c_i} \right)}{k}$$

We present comparative results of clustering on soybean large data and zoo data respectively.

Table 1**:** Cluster recovery for the soybean data with the initial cluster centers computed by the proposed method.

| Clusters found | Objects in cluster | Coming from | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| C1 | 12 | 0 | 12 | 0 | 0 |
| C2 | 15 | 0 | 0 | 15 | 0 |
| C3 | 20 | 20 | 0 | 0 | 0 |
| C4 | 55 | 0 | 0 | 0 | 55 |

Table 2:Comparison of clustering results of traditional and proposed initialization methods on the soybean data.

| The k-modes algorithm | Traditional | Proposed |
|---|---|---|
| AC | 0.8564 | 1.0000 |
| PR | 0.9000 | 1.0000 |
| RE | 0.8402 | 1.0000 |

Table 3: Cluster recovery for the zoo data with the initial cluster centers computed by the proposed method.

| Clusters found | Objects in cluster | Coming from | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| C1 | 10 | 0 | 10 | 0 | 0 |
| C2 | 15 | 0 | 0 | 15 | 0 |
| C3 | 22 | 22 | 0 | 0 | 0 |
| C4 | 50 | 0 | 0 | 0 | 50 |

Table 4: Comparison of clustering results of traditional and proposed initialization methods on the zoo data.

| The k-modes algorithm | Traditional | Proposed |
|---|---|---|

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

| | | |
|---|---|---|
| AC | 0.6447 | 0.8475 |
| PR | 0.7824 | 0.9478 |
| RE | 0.6754 | 0.9147 |

Table 5: comparison of Computational time results of traditional and proposed initialization methods on both soybean and zoo data.

| Categorical Dataset used | Traditional | Proposed |
|---|---|---|
| Soybean Large Dataset | 3.9565 | 2.6755 |
| Zoo Categorical Data | 2.6482 | 1.8647 |

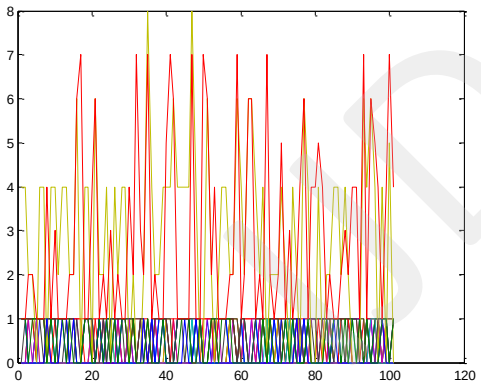


Figure4: Input view of Zoo Categorical data, plotting is done through MATLAB after saving the dataset in form of matrix, different attributes are shown through different color lines

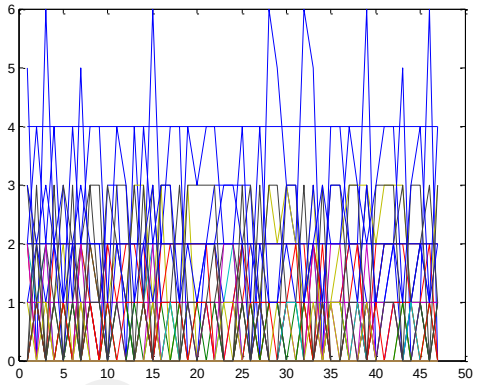


Figure 5: Input view of Soybean Large Categorical data, plotting is done through MATLAB after saving the dataset in form of matrix, different attributes are shown through different color lines

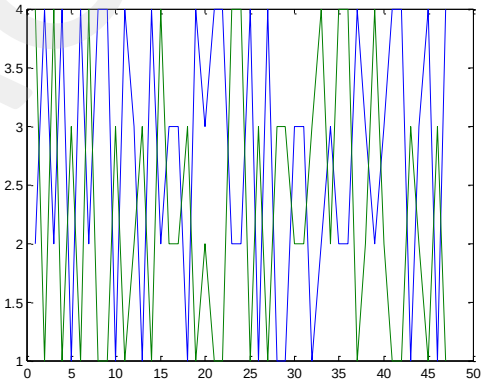


Figure 6: Extracted Cluster membership values from different clusters classified after the application of our genetically optimized fuzzy k-mode algorithm

IJDACR
International Journal Of Digital Application & Contemporary Research

**International Journal of Digital Application & Contemporary research**
**Website: www.ijdacr.com (Volume 2, Issue 1, July 2013**

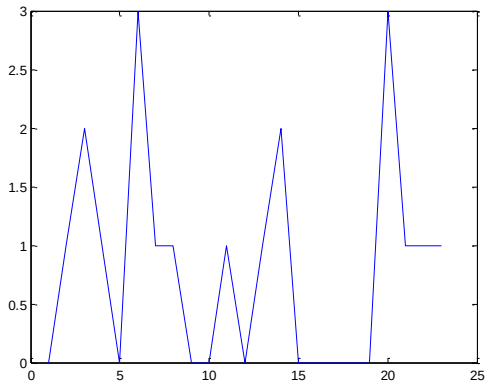


Figure 7: Figure: variation in value of centroid for different clusters at a given instant of time

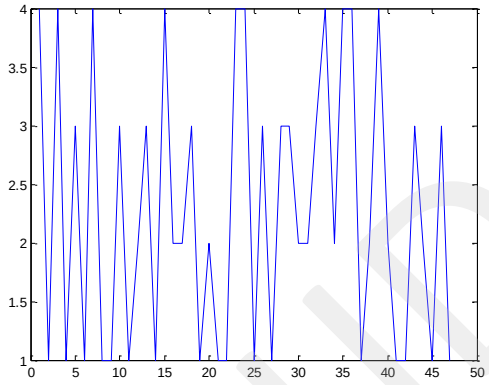


Figure 8: Extracted Attribute variations in classification from different clusters classified after the application of our genetically optimized fuzzy k-mode algorithm

## References

[1] Dr. Shri Kant, Shehroz S Khan, "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation" 2007. IJCAI-papers07/Papers/IJCAI07-447.pdf

[2] S. Aranganayagi and K.Thangavel, "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure". International Journal of Information and Mathematical Sciences 5:2 2009

[3] He, Z., Xu, X. and Deng, S., "A cluster ensemble for clustering categorical data", Information Fusion, Vol. 6, pp. 143-15, 2005.

[4] Ralambondrainy, H.," A conceptual version of the K-Means algorithm", Pattern Recognition Letters, Vol. 16, pp. 1147-1157, 1995.

[5] Gowda, K.C. and Diday, E:, "Symbolic clustering using a new dissimilarity measure", Pattern Recognition Letters. Vol. 24 (6), pp.567-578, 1991.

[6] Huang, Z., "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, Vol. 2, pp. 283-304, 1998

[7] Chaturvedi, A., Green, P., and Carrol, J., " K-modes clustering", Journal of Classification, Vol. 18, pp.35-55, 2001.

[8] Huang, Z., "A note on k-modes clustering", Journal of Classification, Vol. 20, pp. 257-26, 2003.

[9] San, O.M., Huynh, V.N., Nakamori, Y., "An alternative extension of the k-means algorithm for clustering categorical data", International Journal Applied Mathematic Computing Science, Vol. 14 (2), pp. 241-247, 2004.

[10] Liang Bai ,Jiye Liang , Chuangyin Dang, Fuyuan Cao, "A cluster centers initialization method for clustering categorical data". Expert Systems with Applications 39 (2012) 8022–8029