ISSN:0975-9646

Using Genetic Algorithms and Rough Set Based Fuzzy K-Modes to Improve Centroid Model Clustering Performance on Categorical Data

Divyam Sharma , Rishabh Srivastava
Indian Institute of Technology, Kharagpur

Abstract: We propose an algorithm to cluster categorical data named as 'Genetic algorithm initialized rough set based fuzzy K-Modes for categorical data'. We propose an amalgamation of the simple K-modes algorithm, the Rough and Fuzzy set based K-modes and the Genetic Algorithm to form a new algorithm,which we hypothesise, will provide better Centroid Model clustering results, than existing standard algorithms. In the proposed algorithm, the initialization and updation of modes is done by the use of genetic algorithms while the membership values are calculated using the rough set and fuzzy logic.

Keywords:- K modes clustering, Genetic Algorithm, Categorical Data, Rough Sets, Fuzzy logic.

1. INTRODUCTION

Clustering is an unsupervised learning process which divides the entire input data into K regions governed by some similarity/dissimilarity metric. In this case, the number of clusters (K) may or may not be known from before. If we assign each data point to a particular cluster,it is called crisp clustering but in a fuzzy approach, each data point of the data set has a certain degree of membership or membership value corresponding to each cluster.

Centroid model clustering algorithms are iterative in nature and base themselves on the assumption that a data point can be assigned or not assigned to a cluster based on its closeness(of a distance measure) to the central tendency metric for the cluster[1]

Most of the clustering algorithms which have been designed in the last few years are focusing on numerical data. But, many real life datasets have a categorical component[2]. And hence, often these algorithms are rendered ineffective on non-numerical,categorical datasets.

Since there is no built-in distance measure in categorical data, the calculation of the mean of the set of feature vectors cannot be used and hence the standard K-Means algorithm holds no significance. K-Modes algorithm has been in use for some time to deal with categorical data. In this algorithm, the cluster modes are calculated by a frequency based method rather than the mean [3]. Later, the fuzzy versions of this algorithm is also proposed i.e. Genetic Algorithm based Fuzzy K-Modes and Fuzzy K-Modes[4][5].
There is a major disadvantage of using these existing algorithms, which is that they very frequently converge to local optimum solutions. Therefore, these algorithms do not work well for vague, overlapping and uncertain datasets, with many local optimums and where identification of globally optimum clustering poses challenges. Rough Set was introduced as an alternative tool to overcome these difficulties [6].

Rough set is a new concept which when used along with fuzzy set can be applied for fuzzy-rule extraction, fuzzy modelling etc. Both Rough sets and fuzzy sets are known for providing a mathematical base which can take hold of the uncertainties associated with the data. Now-a-days, these algorithms have also made their way into the clustering processes to deal with the vagueness, incompleteness and uncertainty for numerical data sets.

Genetic algorithms(GAs) [7] use the principles of evolution and natural genetics for searching and optimization. GA's are well-known for their ability to perform searching in complex, large and multi-dimensional datasets. They can provide the best possible solution. The algorithm starts by initializing a population of potential solutions encoded into the genes strings called chromosomes. The algorithm proceeds by different processes like computing the fitness values associated with different solution sets. The new generation is created by selection, crossover and mutation. Thus, the correctness of the solution increases as the number of generations increases.

Past research has worked actively in the field of K-Modes clustering [3],[10],[11] and its applications on categorical data. We have seen some people combining Rough sets and fuzzy logic with K-Modes algorithm [5]. Genetic algorithm is also used along with the conventional K-Modes algorithm [4]. Therefore, in the proposed algorithm we have combined Rough sets and fuzzy logic with the Genetic Algorithm and have proposed an algorithm called the Genetic algorithm initialized Rough sets based Fuzzy K-Modes clustering for categorical data.

2. OBJECTIVE

We aim to formulate an algorithm which can be effectively used for clustering text datasets after converting them into categorical format. Our proposed algorithm will merge three algorithms (K-modes, Rough and Fuzzy sets based K-modes and Genetic algorithm) to formulate a new one which we suppose will provide a

higher accuracy that all these existing forms of centroid model clustering algorithms.

Steps for Algorithm Formulation And Application

1. Initiatie the algorithm by randomly assuming the values of some variables like the number of clusters, fuzzifier, weight_low and weight_up.
2. Then, the sum of similarity scores of the first vector with the rest of the vectors is computed and stored in a 1-D array. These values are the fitness values of each vector.(see section 2.3 for details)
3. Repeat Step 2 for all the vectors.
4. A set of cluster centers are selected by randomly selecting some values from the 0^{th} , 1^{st} ,2^{nd} ……n^{th} position of all vectors. The 1^{st} values chosen in all the positions form the 1^{st} cluster center. The 2^{nd} chosen form the 2^{nd} vector and so on. Now a set of vectors, capable of acting as cluster centers is generated.(see section 2.2 for details)
5. Calculate and store the membership values of all the vectors in the clusters represented by these cluster centroids.(see section 2.4 for details)
6. Get a new set of vectors from the data set through distance minimization by selecting the vector with the lowest fitness value (calculated in Step 2) to act as a new set of cluster center provided the membership value of the vector which is to be called the cluster center is greater than 0.01 in that particular cluster. So now, we have a new set of vectors which will be called the new set of cluster centers.(see section 2.5 for details)
7. Now, select 3 centers from the set of cluster centers formed in step 6 as parents by the Roulette wheel selection technique.(see section 2.6 for details)
8. Perform Crossover between these 3 parents by the occurrence based gene selection technique to form one new vector.(see section 2.7 for details)
9. Mutation is done with some probability on this new vector. This step gives us a new vector which will be one of the cluster centers for the next iteration.(see section 2.8 for details)
10. Repeat Steps 7-9 until the required number of cluster centers are not generated.
11. Repeat Steps 5-10 until the criteria for the maximum number of iterations is not fulfilled.(see section
12. Print the Membership values of each vector in each cluster and a graph corresponding to the same.(see section 2.9 for details)

PSEUDO CODE

Input the data

Convert the text data into categorical format

y_true is the actual prime cause

```
sample=[]
modes=[]
for i in range(clusters):
   sample=[]
   modes.append([])
   for j in range(len(vector)):
      if y_true[j]==i:
         include vector[j] in the sample
   transpose sample
   for k in range(len(vector[0])):
      modes[i].append(a random value from sample[k])
distance_sum[]=sum of similarity scores of each vector with all other vectors
while count<50:
   membership[][]=store membership values of each vector in each cluster
   for I in range(len(modes)):
      for j in range(len(vector)):
         modes1.append(vector with minimum distance_sum and membership[l][i]>0.1)
         h.append(the corresponding distance_sum)
   #Selecting the parents by roulette wheel selection technique
  #modes1 is the new set of modes
      Sort h in descending order using bubble sort technique
      h1=[]
      sum1=0
      modes=[]
      for i in range(len(h)):
         add h[i] to sum.
         Append h[i] to h1
      While len(modes)<clusters:
         Parents=[]
         Sum=sum1
         While(len(parents)<3):
               P= a random number between 1 and sum
               Select a parent from modes1 using the roulette wheel selection technique
      #After 3 parents are selected
      Perform crossover between the three using occurrence based gene scanning technique to form one vector
      Perform mutation with some probability
      Append this one vector to modes.
      Count++
Y_pred=[]
Transpose membership
For I in range(len(vector)):
   Y_pred.append(np.max(membership[i])
```

Form a confusion matrix using y_true and y_pred and display the results.

3. PROPOSED GENETIC ALGORITHM AND ROUGH SET BASED FUZZY K-MODES

In this section, we describe the proposed method, called Genetic Algorithm initialized Rough Set Based Fuzzy K-Modes (GARFKMd) which uses the concepts of Rough Sets, Fuzzy Sets and Genetic Algorithm.

3.1 Basic Principle

In this Algorithm, we are using genetic algorithm intuition(along with crossover and mutation) as a way of initializing the initial cluster centroids .The Genetic Algorithm is also used to find the cluster centroid by the distance minimization technique. We also make use of the

Rough and Fuzzy set concept for calculating the membership values of each data point in each cluster.

3.2 Chromosome Representation

Each chromosome has K genes and each gene of the chromosome has an allele value chosen randomly from the set {1, 2, . . . , n}, where K and n denote the number of clusters and the number of points respectively. Hence a chromosome is represented as a vector of indices of the points in the data set. Each point index in a chromosome implies that the corresponding point is a cluster centroid [7].

Example 1. Let K = 5, i.e., Then the chromosome

53 22 95 69 72

represents the indices of five points qualified for cluster centroids. A chromosome is considered as valid only if no point index is present more than once in the chromosome.

3.3 Distance Metric or the Dissimilarity Function

As we know, the inherent distance measuring attributes cannot be used to calculate the distance between two attributes because of the absence of any natural ordering between the data points of the categorical data. In this article, a different distance measure is used. Let us assume two categorical data objects represented by p categorical attributes, $x_i=[x_{i1},x_{i2},\ldots\ldots\ldots\ldots,x_{ip}]$ and $x_j=[x_{j1},x_{j2},\ldots\ldots\ldots\ldots,x_{jp}]$. The distance measure $D(x_i,x_j)$ [3] is defined as the total number of mismatches of the corresponding attribute categories of the objects. That is,

$$D(x_i, x_j)= \sum_{k=1}^{p} \delta\left(x_{ik}, x_{jk}\right)$$

Where,

$$\delta(x_{ik},x_{jk})=\begin{cases}0 & \text{if } x_{ik} = x_{jk}\\ 1 & \text{if } x_{ik} \neq x_{jk}\end{cases}$$

One point to note is that this $D(x_i, x_j)$ gives equal importance to all the categories of an attribute.

3.4 Fuzzy Membership value Initialization

During the initialization stage the K modes selected are denoted by v_l, $1 \le l \le K$.

fuzzy membership value is calculated for all the data objects as follows [5]:

$$\mu_{li}=\begin{cases}1, & \text{if } x_i = v_l\\ 0 & \text{if } x_i = v_h \text{ and } l \neq h\\ \dfrac{1}{\sum_{k=1}^{K}\left[\frac{D(v_l,x_i)}{D(v_h,x_i)}\right]^{\frac{1}{m-1}}} & \text{if } x_i \neq v_l,\ x_i \neq v_h,\ 1 \le h \,\&\, l \le K\end{cases}$$

Here, $D(v_l,x_i)$ is the dissimilarity measure between the cluster mode v_l and object x_i , where $1 \le i \le n$ and n is the number of categorical objects. Here, m is the fuzzy exponent or the fuzzifier.

3.5 Fitness Computation and updating the modes

The fitness computation proposed in this algorithm consists of two parts. In the first phase, all the vectors are assigned their membership values with each mode according to the process mentioned in 2.4.

This step is executed after all the vectors are assigned with their membership values for each cluster. The cluster centroids are replaced by the points having minimum total distance, to the rest of the points, provided their membership in the respective cluster is greater than 1%. In other words, for cluster C_i , the new centroid is the point x_t , where,

$$X_t = \underset{\mu_{ti}\ge 0.01}{\arg\min} \sum_{1\le j\le n} D(x_i, x_j)$$

where, D is the dissimilarity measure.

3.6 Selection

The selection process selects chromosomes directed by the survival of the fittest concept of natural genetic systems [7]. Roulette Wheel selection is one common technique that implements the proportional selection strategy [8]. In the Roulette wheel selection technique, all the chromosomes in the mating pool are arranged in the descending order of their fitness values. Then a random value is generated in between 1 and the sum of all fitness values. We traverse the list of chromosomes arranged in the order of their fitness values and take the sum of the fitness values side by side. That chromosome is chosen as a parent corresponding to which the cumulative sum of the fitness values just exceeds the random value generated earlier. This entire process is repeated 3 times in each iteration to generate 3 parents for crossover.

3.7 Crossover

Crossover is a process that exchanges information between two or more parent chromosomes for generating one or more child chromosomes. In this article, occurrence-based scanning [9] is used to perform the crossover between 3 parents to give 1 child. Occurrence-based scanning (OB-Scan) says that the value whose frequency is the most at a particular position in the parents (which are selected based on the process described in section 2.6) is probably the best possible value to choose. The values are thus chosen by applying a majority function. If the majority function is undecided, then the value is chosen randomly. This occurrence-based scanning technique is applied using the marker update mechanism. An example of how this marker update mechanism works is shown in figure below.

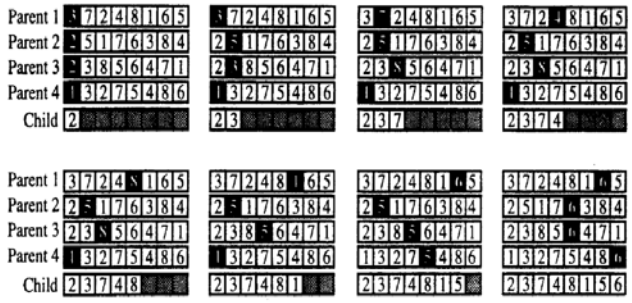

Fig:- OB-scan on order-based representation

3.8 Mutation

Each chromosome is mutated with a fixed probability μ_m [7]. This operation is defined as follows: From the string to be mutated, a random element is chosen and it is inverted (i.e. if the element selected is a 0 it is changed to 1 and vice versa).

3.9 Termination Condition

In this article the algorithm is run for a fixed number of iterations. On termination, we display the membership values of all the vectors in each of the clusters. A plot of the membership values on a graph with the x-axis values representing the cluster number, y-axis values representing the data point number and the membership values are proportional to the intensity of the colour of the point. We also print the confusion matrix by assuming that any data point is a member of the cluster in which it has the maximum membership value.

4. Experimental Results

The proposed algorithm was applied on an accident text data set in which the first column has the description of the accident and the second column contains the primary cause because of which the accident happened. This data set had 709 data points which were in 45 clusters but for the sake of proper visualization and understanding of the readers, the following table and graph is formed using only the first 175 data points.

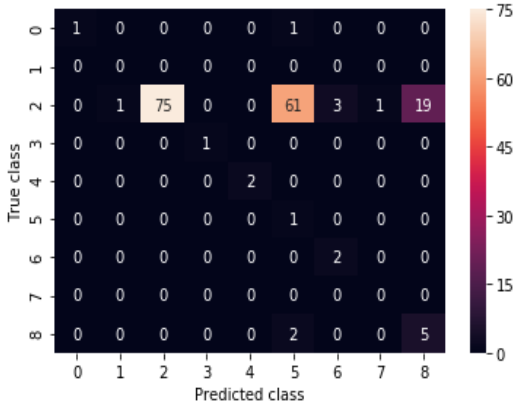

Fig 1:- Confusion Matrix for the proposed Algorithm

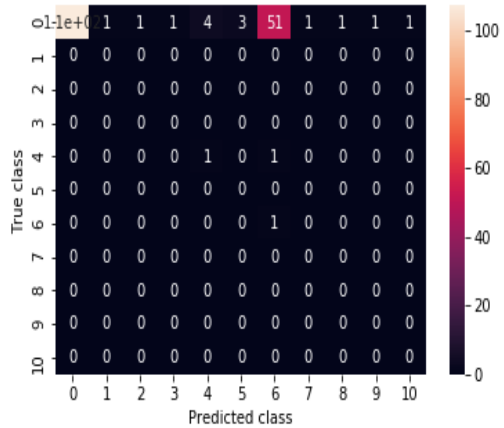

Fig 2: Confusion Matrix when Rough and Fuzzy Set based K modes Algorithm was executed on the same data

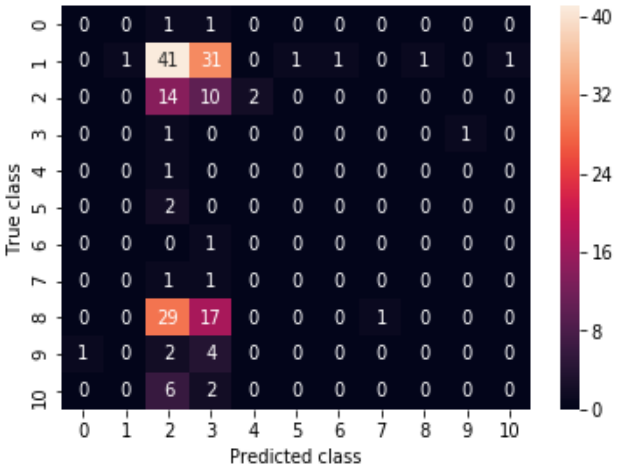

Fig 3: Confusion Matrix when Normal K modes algorithm was executed on the same data

In the above shown confusion matrices, the x-axis shows the predicted labels and the y-axis shows the actual labels. So any box at (x,y) in the confusion matrix shows the number of vectors which are originally in the y^{th} cluster but are predicted to be in the x^{th} cluster.

Algorithm	Precision	Recall	F1-Score
K-Modes	0.1039	0.050135	0.0228
Rough and Fuzzy Set based K-Modes	0.1108	0.193	0.0991
Proposed Algorithm	0.5137	0.52033	0.47028

5. Conclusion

In this paper, Genetic Algorithm and Rough set based fuzzy K modes algorithm is suggested for clustering categorical data. It is seen that the results in terms of F1 scores are enhanced by 4.7 times.This process very effectively optimizes the error function in the normal Fuzzy K-Modes algorithm. The algorithm is tested on an accident data set and the results revealed that this algorithm can be efficiently used for clustering categorical data.

References

1. *Centroid Based Clustering Algorithms- A Clarion Study* Santosh Kumar Uppada / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7309-7313
2. Jinwook Seo & Heather Gordish-Dressman (2007) *Exploratory Data Analysis With Categorical Variables: An Improved Rank-by-Feature Framework and a Case Study, International Journal of Human-Computer Interaction,* 23:3, 287-314, DOI: 10.1080/10447310701702519
3. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,* ZHEXUE HUANG,1998 (ACSys CRC, CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra, ACT 2601, Australia, huang@mip.com.au)
4. Wu, Jianhong & Yang, Z.. (2009). A genetic fuzzy k-Modes algorithm for clustering categorical data.. Expert Syst. Appl.. 36. 1615-1620.

5. Mukhopadhyay, Anirban & Maulik, Ujjwal & Bandyopadhyay, Sanghamitra. (2009). *Multiobjective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes*. Evolutionary Computation, IEEE Transactions on. 13. 991 - 1005. 10.1109/TEVC.2009.2012163.
6. *Rough Set Based Fuzzy K-Modes for Categorical Data by* Indrajit Saha, Jnanendra Prasad Sarkar, and Ujjwal Maulik (Department of Computer Science and Engineering, Jadavpur University, Kolkata - 700032, West Bengal, India indra.raju@gmail.com, jpsarkar@outlook.com, umaulik@cse.jdvu.ac.in),
7. *Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering*, Indrajit Saha and Anirban Mukhopadhyay, January 2009.
8. *Selection methods for genetic algorithms*, Khalid Jebari (Abdelmalek Essaâdi University, December,2014)
9. *Genetic algorithms with multi-parent recombination*, A.E. Eiben, P-E. Rau6, Zs. Ruttkay (Artificial Intelligence Group, Dept. of Mathematics and Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081a 1081HV Amsterdam, 1994)
10. *K Modes Clustering*, Anil Chaturvedi, Kraft Foods, Paul E. Green(The Wharton School, University of Pennsylvania), J. Douglas Carroll (Rutgers University)-2001.
11. *A dissimilarity measure for the k-Modes clustering algorithm*, Fuyuan Cao, Jiye Liang, Deyu Li, Liang Bai, Chuangyin Dang - 2011