A COMBINED APPROACH OF ROCK AND GENETIC ALGORITHM FOR CLUSTERING CATEGORICAL DATA

N. Sowmiya* and B. Valarmathi**

*Research Scholar, School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India.
**Associate Professor and Head of the Department, Software and Systems Engineering, School of Information Technology and Engineering, VIT University, Vellore, Tamil Nadu, India

ABSTRACT: In this paper, a hybrid approach for clustering categorical data is developed. It integrates the ROCK clustering algorithm with genetic algorithm for the betterment of accuracy. The developed approach is tested for four datasets obtained from the University of Irvine (UCI) machine learning repository. Accuracy is used as the measure of performance for accessing the quality of the clusters. The performance of the integrated approach is compared with the performance of ROCK (robust clustering using links) clustering algorithm. The result shows that the integrated approach yields higher accuracies than the ROCK clustering approach.

KEYWORDS: Data Mining; Clustering; ROCK; Genetic Algorithm; Accuracy; Categorical Data;

INTRODUCTION

Data mining techniques like classification and clustering play a vital role in all fields such as education, health care, marketing, industries, etc. Clustering is said to be unsupervised learning because it does not use the predefined class label for clustering. Classification is called as supervised learning because the use of predefined class label. Clustering is the fundamental task in data mining. Its objective is to partition the dataset consists of 'n' objects into 'k' clusters. The objects within the same cluster are similar to each other than the objects in the other clusters (A. K. Jain and Murthy (1999)). There are five types of clustering approaches available namely, hierarchical approach, partition based approach, density based approach, model-based approach and grid based approach.

Many algorithms were proposed for clustering categorical data alone. For example, ROCK proposed by (Guha et al. 2000) and COOLCAT proposed by (Barbará et al. 2002) were the clustering algorithms for clustering categorical datasets. Squeezer is a clustering algorithm and it was proposed by (Zengyou et al. 2002) for categorical attributes. It is more suitable for clustering data streams. This algorithm is suitable for solving small dataset only. For handling of large datasets, they proposed an enhanced algorithm called d-squeezer. Mingoti and Matos (2012) compared some existing categorical clustering algorithms like average linkage clustering, ROCK, *k*-modes, fuzzy *k*-modes and *k*-populations using Monte Carlo simulation. The algorithms like average linkage, ROCK, *k*-modes, fuzzy *k*-modes and *k*-populations were compared. (Dutta et al 2005) proposed a quick version of ROCK algorithm using link graph for clustering categorical data. (Seman et al. 2013) proposed a medoids based clustering algorithm called k-Approximate Modal Haplotype (k-AMH). It was compared with k-modes, k-population, and fuzzy k-modes algorithm. (Seman et al. 2015) enhanced the k-AMH algorithm using the same procedures but with the addition of two methods called new initial center selection and new dominant weighting methods.

Garai and chaudhuri (2004) proposed a two stage novel genetic algorithm for clustering. (Chang et al. 2009) proposed a genetic algorithm for K-means algorithm. (Chang et al. 2012) developed a new genetic algorithm method using message based similarity measure. (Hatamlou 2013) developed an optimization algorithm named Black hole for data clustering. Black hole algorithm also started with the initial population solutions for an optimization problem like other population based methods. In each iteration, the best candidate was selected to the black hole. (Yang et al. 2015) developed a k-modes type clustering algorithm for categorical data which improves the quality of the clusters by using non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC) which combines fuzzy genetic algorithm and multi-objective optimization.

From the literature, it seems that no one has used the proposed hybrid method by integrating ROCK algorithm and a genetic algorithm approach for clustering categorical data.

PROPOSED METHOD

The proposed method consists of two stages. In the first stage, ROCK clustering algorithm is used to find the clusters by varying the threshold value θ from 0.1 to 1 instead of fixing the θ to a certain value. The clusters formed in each iteration acts as an input to the second stage. In the second stage, the genetic algorithm based approach is developed to find the maximum accuracy for the given datasets. The flowchart of the proposed method is shown in Figure 1.

ROCK Algorithm

ROCK is an agglomerative hierarchical clustering algorithm developed by Guha et al. (2000) for clustering categorical data. The steps involved in ROCK algorithm are given below:

Step1: Consider the input dataset consists of 'n' rows and 'm' attributes. Table 1 shows the sample CAR dataset from UCI repository. Totally it has 1728 instances and 6 attributes belong to 4 classes. For the sake of explaining the concept, we have taken 12 instances only (i.e.) 3 instances from each class. The row represents the instances (R1 to R12) and the column represents the attributes (A1 to A6). The last column represents the class label (1 to 4) of the instances.

Table. 1. Sample CAR dataset

	A1	A2	A3	A4	A5	A6	Class
R1	3	3	2	2	0	0	1
R2	3	3	2	2	0	1	1
R3	3	3	2	2	0	2	1
R4	3	1	2	4	0	2	2
R5	3	1	2	4	1	2	2
R6	3	1	2	4	2	1	2
R7	1	0	2	4	1	2	3
R8	1	0	2	4	2	1	3
R9	1	0	2	5	1	2	3
R10	1	1	2	4	2	2	4
R11	1	1	2	5	2	2	4
R12	1	1	3	4	2	2	4

Step2: Calculate the similarity matrix using the Jaccard similarity coefficient. It is calculated using the 'pdist' function available in MATLAB software. In general, the formula for calculating the Jaccard similarity for the two sets X and Y is shown in equation 1.

$$S(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \qquad (1)$$

Step 3: Compute adjacency matrix 'adj' using the following conditions. The threshold θ is varying from 0.1 to 1. Initially start with θ = 0.1.

```
if $S(X, Y) > \theta$
        $S(X, Y) = 1$;
else
        $S(X, Y) = 0$;
end
```

Step 4: Formation of link matrix by multiplying the adjacency matrix (adjm) by itself to find the number of links. It is shown in equation 2.

$$L = [adjm] x [adjm] \qquad (2)$$

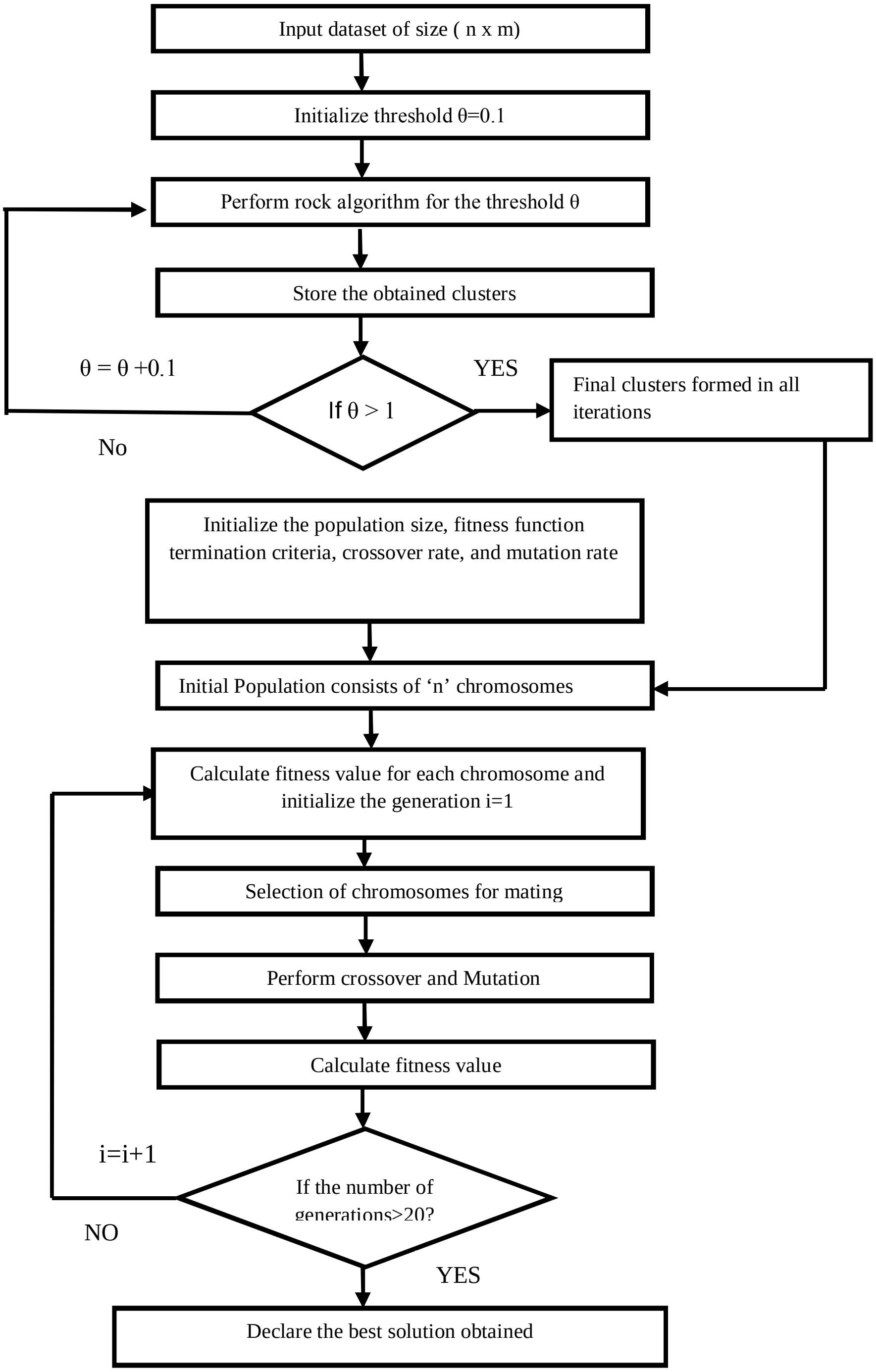

Input dataset of size (n x m)
Initialize threshold θ=0.1
Perform rock algorithm for the threshold θ
Store the obtained clusters
θ = θ +0.1
If θ > 1
YES
No
Final clusters formed in all iterations
Initialize the population size, fitness function termination criteria, crossover rate, and mutation rate
Initial Population consists of ‘n’ chromosomes
Calculate fitness value for each chromosome and initialize the generation i=1
Selection of chromosomes for mating
Perform crossover and Mutation
Calculate fitness value
i=i+1
If the number of generations>20?
NO
YES
Declare the best solution obtained

Figure. 1. Flowchart of the proposed method

Step 5: Calculation of goodness value for the merging clusters g(C$_i$, C$_j$) is shown in equation 3.

$$g(C_i, C_j) = \frac{L[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \qquad (3)$$

Where,

- C$_i$, C$_j$ represents the cluster i and j respectively
- g(C$_i$, C$_j$) is the goodness value for the cluster i and j
- L[C$_i$,C$_j$] is the link value for the clusters i and j
- n$_i$= number of points in C$_i$

- n_j= number of points in C_j
- θ represents the threshold value $f(\theta)$ =[(1-θ) / (1+θ)]

Step 6: Merge two points which are having maximum goodness value. Then modify the link matrix and repeat step 5 until all elements are arranged in the clusters. If we need any specific 'k' clusters it will automatically terminate after reaching the 'k' clusters.
Step 7: The final clusters formed for each theta value are stored.
Step 8: Increment the θ value by 0.1 and repeat the steps 3-7 until θ value is reached to 1.

Genetic Algorithm based approach

A computational form of the Genetic algorithm (GA) was introduced by (Holland 1975). It is the meta-heuristics method used to find the solution for the optimization problems and search problems using the iterative search procedure. GA is the subclass of evolutionary algorithms (EA). It depicts the process of natural evolution principles such as selection or reproduction and natural genetics. It starts with the set of chromosomes or otherwise named as the initial set of solutions called initial population. Each chromosome contains the fixed length of genes of strings of any data type. It adopts the concept called Darwin's "Survival of the fittest".

The steps involved in GA are given below:
Step 1: Formulation of fitness function
The fitness function is nothing but the objective function for the chosen problem. It may be either maximization or minimization of the objective function. It consists of any number of variables. Here the fitness function value (F) is the maximization of accuracy is given in equation 4 and the formula for calculating the accuracy is given in equation 5.

$$F = \max(\text{Accuracy}) \qquad (4)$$

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}} \qquad (5)$$

Step 2: Encoding of chromosome
Each chromosome contains the fixed length of genes of any data type. The length of the chromosome chosen is 12 for the sample dataset. It is the cluster number of instance where it actually belongs to. Figure 2 depicts the encoding of chromosome, where the numbers 1, 2, 3, 4 represents the class label. From Figure 2, the number 1 present in first three positions represents that the instances R1, R2, R3 are grouped in cluster1. The value 2 present in the fourth position represents R4 belongs to cluster2, and so on.

1	1	1	2	3	4	2	2	1	3	4	2

Figure. 2. Representation of chromosome

The genes in the chromosome are generated either randomly or by using some procedures. Here the chromosomes are generated from ROCK algorithm by varying θ values.

Step 3: Initialization
In this step, initialize the size of population, crossover & mutation rates, and termination criteria. The population size is of any size depends on the user. Here the population size chosen is 6 because the ROCK algorithm produces only 6 chromosomes for the dataset shown in Table 1. It means that the clusters are formed up to θ=0.6 after θ=0.6 all the values in the adjacency matrix becomes zero so no clusters are formed. The crossover rate P_c may be of any random value between 0.1 and 1. Similarly the mutation rate P_m lies between 0.001 and 0.01.The termination criteria chosen here is the specific number of iterations. Table 2 shows the generated chromosomes obtained in stage 1 for the dataset shown in Table 1 using the ROCK algorithm.

Step 4: Evaluation
Compute the fitness function value for the generated chromosome using equation 4 and 5.

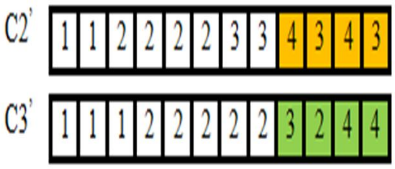
Step 5: Selection
Many methods are available for selection operation. Here Roulette wheel selection is used for selection of chromosomes for the next generation. The chromosomes selected for crossover are C5, C2, C6, C1, C3, and C3. The selected chromosomes are named as C1' to C6'.

Table. 2. Initial Population

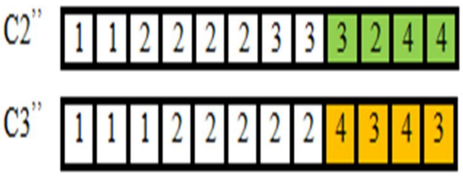
Chromosomes												
C1	1	1	1	2	3	4	2	2	1	3	4	2
C2	1	1	2	2	2	2	3	3	4	3	4	3
C3	1	1	2	3	2	2	3	3	4	3	3	3
C4	1	1	2	3	3	4	3	4	4	4	4	4
C5	1	2	3	4	4	4	4	4	4	4	4	4
C6	1	1	1	2	2	2	2	2	3	2	4	4

Step 6: Crossover

Crossover operator is used to randomly exchange the genes of the pair of chromosomes to create two new offspring. The random position chosen is 9 and exchanges the genes after that position to create a new offspring. Figure 3 and 4 represent the crossover operation.



C2'	1	1	2	2	2	2	3	3	4	3	4	3
C3'	1	1	1	2	2	2	2	2	3	2	4	4

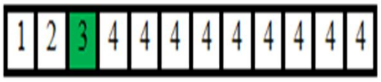


C2''	1	1	2	2	2	2	3	3	3	2	4	4
C3''	1	1	1	2	2	2	2	2	4	3	4	3

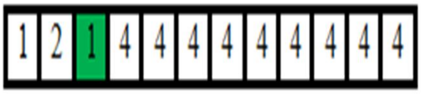
Figure. 3. Before crossover-parents **Figure. 4.** After crossover-offspring's

Step7: Mutation

Each chromosome need to undergo mutation. Perform a bit flip mutation for each chromosome. For example consider chromose 2, the position to be changed is highlited and it is shown in Figure 5. If it is the binary string we can flip that to '1' if the value is '0' and vice versa. Since it is the numeric string change the gene in the chromosome by choosing the value between the number of classes. It is shown in Figure 6.



1	2	3	4	4	4	4	4	4	4	4	4



1	2	1	4	4	4	4	4	4	4	4	4

Figure. 5. Before mutation **Figure. 6.** After mutation

Step 8: At this stage calculate the fitness value for the chromosomes after mutation and consider this as the population for the next generation and perform the steps 4 to 7 until termination criteria is satisfied. The program terminates after 20 iteration.

RESULT AND DISCUSSION

The proposed method is tested for 4 benchmark datasets namely CAR (1728 Instances x 6Attributes→4classes), HAYES (132 Instances x 4 Attributes →3classes), BALANCE SCALE (625 Instances x4→3 classes), BREAST CANCER (699 Instances x 9 Attributes →2 classes), taken from the UCI repositories. The accuracy of CAR dataset is equal in case of both ROCK and the proposed method. For HAYES dataset the accuracy of proposed method is superior than ROCK in terms of 7.40 %. The accuracy of proposed method for BALANCE SCALE dataset is superior than ROCK in terms of 40.68 %. Similarly the CANCER Dataset produces 34.84 % superior accuracy for proposed method than ROCK. The bar chart for the results is shown in Figure 7.

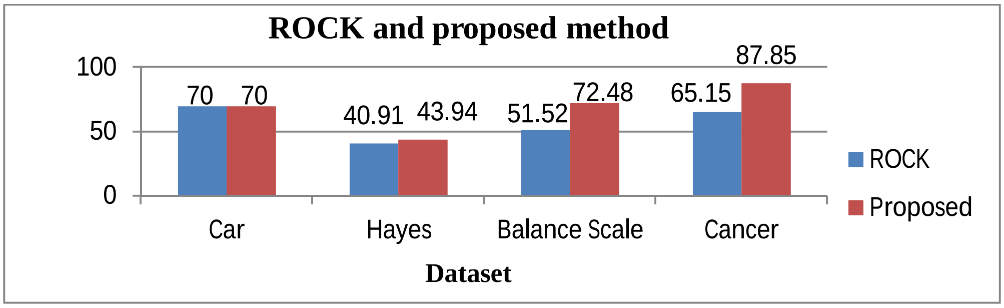


Figure. 7. Comparative analysis of ROCK and proposed method

CONCLUSION

The integrated approach of ROCK and Genetic algorithm for clustering categorical data is developed and the results were compared with the ROCK algorithm for 4 UCI datasets. The comparative results shows that it produces equal accuracy for 25% of the datasets and produces higher accuracy for 75% of the datasets.

REFERENCES

[1] Barbará D; Couto J; Li Y; (2002); *COOLCAT: An entropy-based algorithm for categorical clustering*; In: Proceedings of 11th ACM international conference on information and knowledge management; McLean; VA; USA; 582-589.

[2] Chang D; Zhao Y; Zheng C; Zhang X; (2012); *A genetic algorithm using a message-based similarity measure*; Exp. sys. appl.; **39**; 2194-2202.

[3] Chang DX; Zhang XD; Zheng CW; (2009); *A genetic algorithm with gene rearrangement for K-means clustering*; Pattern Recogn.; **42**; 1210 – 1222.

[4] Darwin CR; *On the Origin of Species by means of Natural Selection and The Descent of Man and Selection in Relation to Sex*; third ed., Vol. 49 of Great Books of the Western World; Editor in chief: M. J. Adler. Robert P. Gwinn; Chicago IL; 1991. First edition John Murray, London; 1859.

[5] Dutta M; Mahanta AK; Pujari AK ;(2005); *QROCK: a quick version of the ROCK algorithm for clustering of categorical data*; Pattern Recogn. Letters; **26**; 2364–2373.

[6] Garai G and chaudhuri B B; (2004); A *novel genetic algorithm for automatic clustering*; Pattern Recogn. Letters; **25**; 173–187.

[7] Guha S; Rastogi R; Shim K; (2000); *ROCK: A Robust Clustering Algorithm for Categorical Attributes*; Inform. Syst.; **25**; 345–366

[8] Hatamlou A; (2013); *Black hole: A new heuristic optimization approach for data clustering*. Info. Sci. **222**; 175–184.

[9] Holland JH; (1975); *Adaptation in natural and artificial systems*. Univ. of Michigan Press, Ann Arbor. USA.

[10] Jain AK; Murthy MN; Flynn PJ (1999); *Data clustering: A review*; ACM Comput. Surv.; **30(3)**; 264-323.

[11] Mingoti SA; Matos R; (2012); *Clustering Algorithms for Categorical Data: A Monte Carlo Study*. Int. J of Stat. and Appl.; **2**; 24-32.

[12] Seman A; Abu Bakar Z; Mohd. Sapawi A; Othman I R; (2013); *A medoids-based method for clustering categorical data*; J. Artificial intelligence; **6**; 257-265.

[13] Seman A; Sapawi AM; Salleh MZ; (2015); *Performance Evaluations of κ-Approximate Modal Haplotype Type Algorithms for Clustering Categorical Data*; Research J. Info. Tech.; **7**; 112–120.

[14] Yang CL; Kuo RJ; Chien CH; Quyen NTP; (2015); *Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering*; App. Soft Comput.; **30**; 113–122.

[15] Zengyou H; Xiaofei X; Shengchun D; (2002); *Squeezer: An Efficient Algorithm for Clustering Categorical Data*; J. Comput. Sci. & Technol; **17**; 611-624.