A Survey Of K-Means And GA-KM The Hybrid Clustering Algorithm

Yogita chauhan, Vaibhav Chaurasia, Chetan Agarwal

Abstract: In this paper we present application of hybrid clustering algorithm that combines partitioning clustering algorithm and heuristic search algorithm. Our method uses portioning method with Genetic algorithm . First we cluster the data using K-Means clustering algorithm with the value of K no. of clusters then we calculate the centroid of K cluster obtain from the previous step. Than we apply Genetic algorithm for centroids for the given value K clusters (GAKM). After applying the GAKM we compare the result of simple K-Means and GAKM algorithm .our experimental results shows that the cluster obtained from GAKM are provides more optimal result in comparison of simple K-Means algorithm cluster result**.**

Index Terms: Evolutionary data mining, clustering, Partitioning methods, K-Means clustering algorithm, Genetic algorithm,, GAKM algorithm, biomedical data,

1 INTRODUCTION

Clustering is a unsupervised classification technique . we can classify the data without using class label by using clustering, and by using the distance measurement we group the cluster set. The cluster is a collection of homogeneous data types many clustering methods are there in data mining in this paper we are using K-Means (partitioning method), where the number of clusters K are given. In K-Means algorithm we divide or partition whole data in to number of K clusters. Each data object belongs to only one cluster. Each cluster must contain at least one data object , these are the basic condition for the partitioning clustering method. K-Means is one of the partitioning method which is applicable for numeric data. When the no. of cluster K is known a prior clustering may be formulated as distribution of n objects in N dimensional space among K groups in such a way that objects in the same cluster are more similar in some aspects than the others in different clusters. This involves minimization of some optimization criterion [1]. Genetic algorithm is purposed by Holland in1975in the university of Michigan. The principle of natural genetics is to construct search and optimization procedures.This is a heuristic search algorithm whose collective behaviour emerges from a group of social insects such as ants, bees, and wasps has been known as Swarm Intelligence. The foraging of ants has led to a novel algorithm called Ant Colony Optimization (ACO) [3].

- *Yogita Chauhan1 is currently pursuing masters degree program in software engineering in Galgotias University, Greater Noida ,U.P, India, PH-07838735230. E-mail: yogitachauhan8@gmail.com*
- *VaibhavChaurasia2 is currently pursuing masters degree program in software engineering in Galgotias University Greater Noida U.P, India, PH-09871457822. E-mail: vaicha.oracle@gmail.com*

The simple Genetic algorithm automatically obtain the knowledge about the search space which is the space for all possible and feasible solutions.GAs are inspired by Darwinian theory of the survival of the fittest. Algorithm is started with a set of solutions (represented by chromosomes) called populations. Solutions for population are taken and used to form a new population. This is motivate by a hope that new population (offspring) , are selected according to their fitness. The more suitable they are the more chance they have to reproduce. This is repeat This is repeat until some conditions(number of populations) for improvement of best solution are satisfied [3].for detailed study of GA readers are referred to [9]. GAKM by Jenn-Long Liu , yu-Tzu Hsu and Chih-Lung Hung is a hybrid approach which combines the Genetic algorithm(GA) and K-means(KM) the Genetic algorithm steps are combined with K-means where the result of K-means is used for setting the objective function of GA. If satisfied condition obtain from here then the best solution is taken out otherwise the GA steps (reproduction, crossover, mutation) are apply to the parameters to calculate the fitness value and get a best solution. This algorithm applied to the biomedical data base (cardiac disease dataset sample) from UCI machine learning repository.It contains 270 instances belonging to two classes - normal (150) and heart patient (120). Each record in the cardiac disease dataset is characterized by 14 attributes, including 13 condition attributes and 1 decision attribute which is the presence of cardiac disease[4]. The next section of this paper contains overview and explanation about K-means algorithm and GAKM algorithm in section 2, 3.

2K-MEANS ALGORITHM

Partitioning based method creates K partitions, called clusters (group of similar data objects) from the given set of n data objects. Initially some data objects are assign to some of the partitions, An alternative relocation technique is used to improve the partitioning by moving objects from one cluster to another cluster. The data objects in the one cluster are similar to each other or we can say them homogeneous data objects, but these data objects are different from the another cluster data objects. The each partition is represented by either a centroid or a medoid, A centroid is the average of the all data objects in a partition, while a medoid is the most representative point of a cluster [8].The fundamental requirements of the partitioning based methods are each cluster must contained at least one data object, and each data

object must belong to exactly one cluster. In this category of clustering the k-means is the simplest and easy method to cluster the data. The similarity measurement for this method is carried out by distance measurement. It is most common to calculate the dissimilarity between two patterns using a distance measure define on the feature space. The most popular metric for distance measurement is Euclidean distance.

$$d_2(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_i, k - x_j, k)^2} = \| x_i, x_j \|^2$$

A flow chart for K-means algorithm is given below

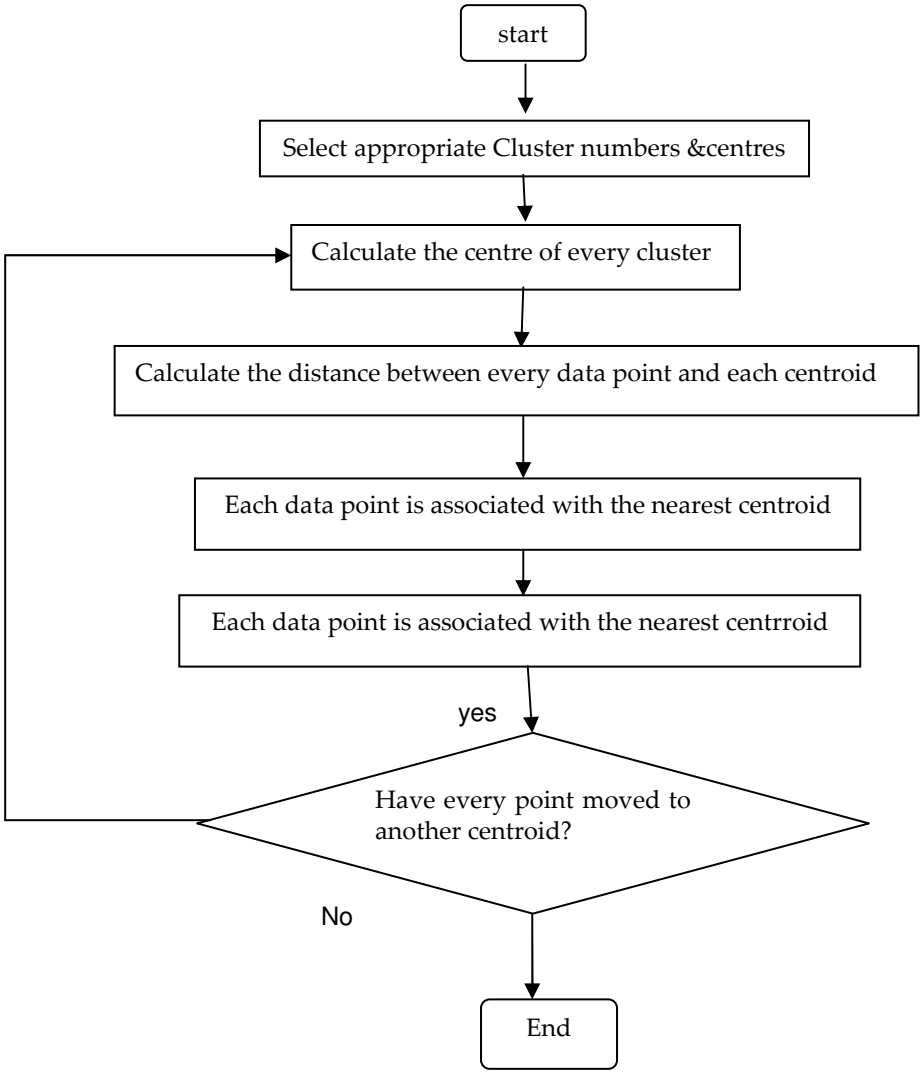

Fig 1 Flowchart of K-means algorithm

An algorithm for K-means partitioning is given below [2]:

Input: 'k', the number of cluster to be partitioned;

'n', the number of objects.

Output: A set of 'k', cluster based on given similarity function

Steps:

i) Arbitrarily choose 'k' objects ad the initial clusters;

ii) Repeat ,

a. (Re)assign each object to the cluster to which the object is the most similar, based on the given similarity function ;

b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster ;

iii) Until no change.

The advantage of K-means algorithm is that it is very efficient and it can be apply for high dimensional data. Some drawbacks are also there for K-means like it is only used when data objects are numeric value or we can say that K-means only applicable for the numeric data. The number of clusters (the k value) is to be define bye user. One of the most important drawback is if outlier and noise point are available

than the mean values are changes, k-means is very sensitive for noise or outlier point because it is going to effect the mean point of the cluster.

3 GENETICALGORITHM – K-MEANS (GAKM)

Jenn-Long Liu, Yu-Tzu Hsu and Chih-Lung Hung [4] proposed GAKM a hybrid method that combines a genetic algorithm (GA) and K-means algorithm. The function of GAKM is to determine the optimal weights of the attributes and cluster centres of clusters that are needed to classify the dataset. Genetic algorithm is a stochastic search algorithm which is based on the Darwinian principal of natural selection and natural genetics. The selection is biased toward more highly fit individuals, so the average fitness of the population to improve from one generation to the next. In general GA generates an optimal solution by means of using reproduction, crossover, and mutation operators [9, 10]. The genetic algorithm initially start with population generated, population is the collection of chromosomes, chromosome is the collection of genes, the fitness for the population is calculated by using a suitable fitness function accordingly. In GAKM the result of K-means algorithm is used for setting the objective function of GA. If fitness value is satisfied, the best solution is obtain. Otherwise the GA parameters (reproduction, crossover, mutation) are apply for obtain a optimal no. of cluster.The flowchart for GAKM is as follows:

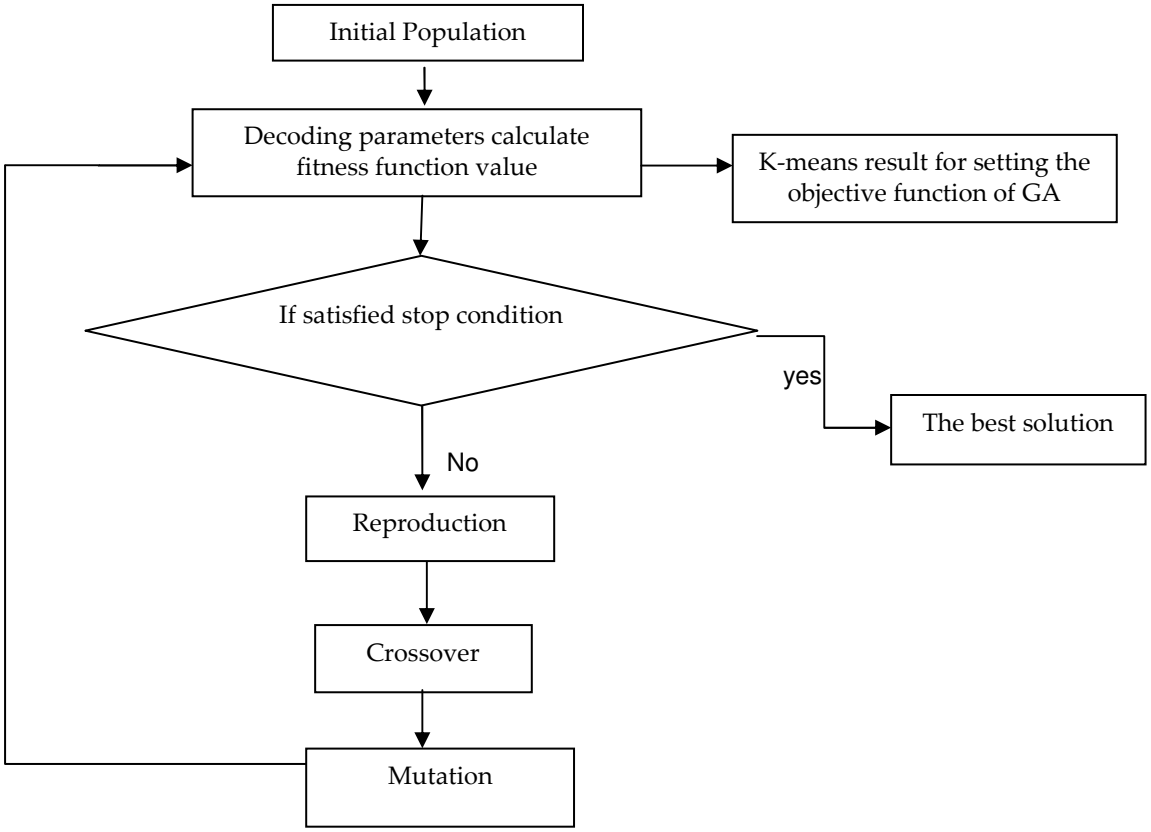

Fig 2 Flowchart of GAKM algorithm

The advantage of GAKM algorithm is that it can apply to high dimensional data. It gives an optimal solution which is the best solution to the problem. The fitness of best individual is also expected to improve over time, and the best individual may be selected as a solution after several generations.

An algorithm for GAKM is give below [11]:

Input:
Mutation probability, Pm;
Population Size, N;
Maximum number if generations, MAX-GEN;

Output:
Solution string, s*;
{ Initialize the population, P;
geno = MAX-GEN;
s* = P1; (Pi is the length in P)
While (geno> 0)
s* = P1; (Pi is the ith string in Pi)
P = Selection (P);
for i = 1 to N, Pi = Mutation (Pi);
for I = 1 to N, K-means (Pi);
S = string in P such that the corresponding weight matrix Ws has the minimum SE measures;
If (S(Ws)) > S(Ws)), s* = S;
Geno = geno-1;
}
Output s*;
}

The steps involve in GAKM are coding, initialization, selection, mutation. The coding is string-of group-numbers encoding [12]. The initial population P(0) is selected randomly. For the selection the roulette wheel strategy is used for random

selection. The use of one step K-Means in GAKM instead of the crossover operator used in conventional GA. In the GAKM biased mutation operator is define specific to clustering called distance based mutation. Thus GAKM combines the simple-Means and GA. K. Krishna and M. NarasimhaMurty conduct experiments to analyse the significance of the operators used inGAKM and the performance of GAKM on different data sets and varying sizes of search spaces.

4 CONCLUSION

In this paper we study the two clustering algorithms one is simple K-Means partitioning algorithm and the GAKM an hybrid algorithm which is the combination of simple K-Means and Genetic Algorithm. K-means is combine with GA to get the optimize no. of clusters from the result of simple K-Means algorithm .Both algorithm are simple to understand and can be applicable for various type of data like genomic data set, numerical data set. Its review from study that K-means applicable only when mean of cluster is defined, it is not applicable on categorical data. But it is easy to understand and implement. GAKM is good for complex problems it retains best features. It is a outcome that the accuracy and performance of GAKM is better than simple K-means.

REFERENCES

[1] TeherehHassanzadeh and mohamad Reza Meybodi ,"A New Hybrid Approach for Data Clustering using Fairfly Algorithm and K-means", the 16th CSI International Symposium on Artificial Intelligence and Signal Processing, ASIP,2012

[2] H. Jiawai and K. Micneline, Data Mining: concepts and techniques, Morgan Kaufmann Publishers, 2001.

[3] S. Rajasekaran and G.A. VijayalakshmiPai, Neural Network, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications, PHI, July 2012.

[4] Jenn-Long Liu, Yu-Tzu Hsu, Chih-Lung Hung , " Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis", WCCI 2012 IEEE World Congress on Computational Intelligence, June 2012.

[5] R. Bhawani, G. SudhaSadasivam, RadhikaKumaran, " A Novel Parallel Hybrid K-means –DE-ACO Clustering Approach for Genomic clustering using MapReduce", World Congress on Information and Communication Technologies, 2011.

[6] Guangya LIU, Jingli CHEN, "The Application of Genetic Algorithm based on Mat lab in Function Optimization",

[7] Shalini S singh, N C Chauhan, "K- means v/s K-medoids; A Comparative study", National Conference on Recent Trends in Engineering & Technology, B. V. M. Engineering College, V. V Nagar, Gujarat, India,May 2011.

[8] T, Velmurugan and T. Santhanam,"A Survey of Partition Based Clustering Algorithm in Data Mining: An experimental Approach", An Experimental Approach. Informational Technology Journal, Vol, 10, No . 3, pp478-484, 2011

[9] D. Goldberg, Genetic Algorithm in Search , Optimization and Machine Learning, Addison Wesley, 1989.

[10] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, 3rd ed., Springer-Verlag, 1999.

[11] K. Krishna and M NarasimhaMurty,"Genetics K-Means Algorithm," IEEE transactions on system ,man and cybernetics, Part-B: Cybernetics , Vol. 29, No. 3, pp. 433-439, June 1999.

[12] R. Jones and M.A. Beltramo, " Solving partitioning problems with genetic algorithms,"in proc. 4th int. conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman, 1991.

Author(s) Profile

YogitaChauhan is pursuing M.Tech. in Software Engineering from Galgotias University Greater Noida, Delhi-NCR. She is completed B.Tech. in Information Technology from Gold Field Institute of Technology & Management, Faridabad affiliated from MaharshiDayanand University, Rohtak.Area of interest are Software Testing, Data Mining

VaibhavChaurasia pursuing M.Tech. in Software Engineering from Galgotias University. He is completed B.Tech. in Computer Science from NIMS University in 2012 and pursuing M.Tech. in Software Engineering from Galgotias University currently in 2014 respectively. Area of interest is Software Testing.

ChetanAgarwal, is an Assistant Professor at Galgotias University, Greater Noida, Delhi-NCR. He has been a student of B.E fromRajsthan University, andM.Tech fromDhiruBhaiAmbani Institute of Information Communication Technology, GandhiNagar (Gujrat),India. He has more than 10years of experience in Teaching. He has taught subjects DBMS, DMS, TOC, DAA, DSA. He has involved in various academic activities. He has wide research interests that include Natural Language Processing, Speech Synthesis & Reorganization, Parsing