INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A COMPREHENSIVE STUDY OF GA BASED CLUSTERING ALGORITHM IN DATA MINING

[1]JaskaranjitKaur, [2]NavneetKaur, [3]JaskiranKaur

[1]Assistant Professor, Lyallpur Khalsa College, Jalandhar, jaskaranjitkaur11@gmail.com
[2]Assistant Professor, Lyallpur Khalsa College, Jalandhar, saininavneet@gmail.com
[3]Assistant Professor, Lyallpur Khalsa College, Jalandhar, Jaskiran.k89@gmail.com

Abstract: - This paper presents a review on GA based clustering Algorithm in Data Mining. Clustering is a practical unsupervised data mining technique that segregates an input data set into a required number of subgroups. K-means is a usually used for partitioning based clustering technique that identifies the user selected clusters (k), identified by their centroids, by minimizing the square error function. Genetic algorithm one of the usually used evolutionary algorithms, performs the exhaustive exploration to discover the result to a clustering problem. A comparative study has been done among GA Based Clustering, K-Means Based Clustering, and Fuzzy C Means Clustering.

Keywords: Data Mining, Clustering, Genetic Algorithms, K-Means

I. INTRODUCTION

Clustering is one of the basic techniques of data mining where similar characteristics of data are grouped. Data Mining is a mining process of knowledge. It is an interdisciplinary regime of computer science [1]. It is the enumerated procedure of observing patterns in immense data sets jumbled with artificial intelligence, machine tutoring, statistics and database systems. Clustering is one of the remarkable data mining algorithms, used for a lot of practical application in various emerging areas like Bioinformatics. The data on the same cluster shares analogous patterns. The clustering overcomes such classification by adapting the changes and the unknown features of the dataset. These days, researchers ponder on clustering owing to its properties such as scalability, skill to deal with different kinds of attributes, discovery of clusters with attribute shape, high dimensionality, ability to deal with noisy data, interpretability[2].The excellence of a clustering result is determined by the likeliness and its execution and also by its potential to explore few or entire of the concealed behavior [3].A range of clustering techniques involved in data mining are center-based, density-based, and conceptual-based clustering methods. Conceptual clustering is a machine learning prototype for classification which distinguishes ordinary data by producing a concept of

description for each generated classes. Most conceptual clustering methods are capable of generating hierarchical class structures. Conceptual clustering is closely related to formal concept scrutiny, decision tree learning, and mixture model learning. In density based clustering, clusters are defined as areas of higher solidity than the remainder of the data set. The objects in these sparse areas require separate clusters and usually are considered to be noisy and border points. As a result a cluster will concise of all density-connected objects beside all objects that are within the objects range. Subsequently, center-based **clustering** represents clusters as a central vector, that might not necessarily be a member of the data set. Genetic Algorithm (GA) parallel search method, that stalks for a universal result for the clustering agitation via the operation of the postulates of natural selection [10]. Several non-GA-based clustering algorithms have been generally used, like K-means, Fuzzy-c-means, EM, etc. Though, the number of clusters in a data set is not known in almost all real-life situations. None of these non-GA-based clustering algorithms is competent of efficiently and automatically forming natural groups from all the input model, especially when the number of clusters included in the data set tends to be large. on the other hand Genetic algorithm directed by the Darwin's theory of evolution and natural genetics portrays the features like mutation, crossover, and the fitness evaluation, which could make it as a powerful tool to optimize the clustering techniques. Yet an additional characteristic of Genetic algorithm is population generation and maintenance of the generated population. The adaption of suitable feature in diversity of the population will be preserved [10].The algorithms instigate from random selected solution called the population and create new successive, fresh propagation of the population by using these genetic manipulators. Natural selection is performed on the fitness of an individual. The more it is fit, more it has the opportunity to persist in the successive genesis. Crossover is achieved by swapping the components by 1point or 2point crossover rule and mutation is projected to alter the string either 0 to 1 or 1 to 0 by the user-specified arbitrary positions. The insight fundamental procedure is that every reborn populated is superior to the earlier one. Usually, the result is illustrated by employing the specific portion of the strings, especially, the binary strings, but voluntary encodings are being evolved. The benefit of genetic algorithms is that the fitness operation is tailored to innovate the performance of the algorithm. In this paper we have discussed Clustering technique which includes partitioning clustering and further we have discussed Genetic Algorithm. In section 4, working of GA based clustering algorithm has been elaborated. Finally a comparison between GA Based Clustering, K-Means Based Clustering, Fuzzy C Means Clustering has been done.

II. CLUSTERING

The process of grouping a set of objects into similar classes is called clustering, as in Clustering can be defined into two categories: Hard clustering and Fuzzy clustering. Hard clustering assign each feature vector to one and only one of the clusters. It must have a degree of membership equal to one and well defined boundaries between clusters. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and fuzzy boundaries between clusters. A cluster is a group of data objects that are similar to one another within the same cluster which means high intra cluster similarity and are dissimilar to the objects in other clusters which means low inter cluster similarity. A cluster of data members can be treated collectively as one group and so many be considered as a form of data compression. Although classification is an efficient means for separating groups or classes of objects, but it is costly and the labeling of a large set of training instances or patterns, which is used by classifier to model each group. It is often more desirable to proceed in the reverse direction, as in First partition the groups are created on the basis of data similarity, and then labels are assigned to the small number of groups. One more advantage of this clustering-based process is that it is easy adaptable to changes and helps find out useful characteristics that distinguish different groups. [4]

Figure 1: Various stages of clustering

A. Partitioning Clustering

A partitioning clustering is a division of large set of data objects into disjoint clusters such that each data object is in one subset. Partitioning clustering decomposes a data set into a set of non-overlapping clusters. Given a data set of *N* points, a partitioning method creates *K* ($N \geq K$) partitions of the data, with each partition depicting a cluster which means it divides the data into *K* clusters by satisfying the following conditions:

(1) Each group contains at least one point, and

(2) Each point belongs to exactly one group.

The main purpose of partitioned clustering algorithms is to minimize an objective function. For example, in *K*-means and *K*-medoids the function also called distortion function is

$$\sum_{i=1}^{k}\sum_{j=1}^{C_i} \boldsymbol{Dist}(\boldsymbol{x_j}, \boldsymbol{center}(\boldsymbol{i})) \qquad \textbf{(1)}$$

Where $|C_i|$ is the number of points in cluster *i*, Dist(x_j, center(*i*)) is the distance between point x_j and center *i*. Many functions for calculating the distance can be used, such as Euclidean distance.

B. K-Means Clustering

The term "k-means" was first used by James MacQueen in 1967.In data mining, k-means clustering is a process of cluster analysis whose main objective is to partition **n** observations into **k** clusters in which each observation belongs to the nearest mean. The result of k-means is such that the resulting cluster should have high intra cluster and low inter cluster similarity. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid [6]

The basic k-means algorithm is as follow:

1. Select k data points as the initial centroids.
2. (Re) Assign all points to their closet centroids.
3. Re compute the centroid of each newly build
Cluster.
4. Repeat step 2 and 3 until the centroids do not
Change.

The k-means algorithm is attractive in practice because it is simple and it is generally fast.

III. GENETIC ALGORITHMS

Genetic algorithms are non-deterministic search algorithms which are based on the mechanics of natural selection and natural genetics in a biological system. The genetic algorithms attempts to find solution to the problem by genetically breeding the population of chromosomes. The genetic algorithm transforms a population of individual chromosomes, each with an associated fitness value, into a new generation. This is based on the theory of Darwinian principle of survival of the fittest and naturally occurring genetic operations such as reproduction, crossover and mutation.

Before applying genetic algorithms, we define a relevant encoding of chromosome (binary or real valued) to solve a problem and also design fitness function. We generate an initial population consisting of chromosomes and evaluate the fitness value of each chromosomes and then we will select two chromosome randomly and crossover and mutate them and finally replace a low quality chromosome with a new chromosome of better fitness. As these processes have been repeated, the population consists of high quality chromosomes. **[**7] [8]

IV.GA BASED CLUSTERING ALGORITHM

The techniques typically starts with a set of randomly generated individuals called the population and procreate successive, latest generations of the population by genetic operations such as natural selection, crossover, and mutation. Genetic Algorithm is used for solving optimization problems. As clustering problems is defined as an optimization problem, so GA is appropriate here. Firstly clusters are built using K Means clustering algorithm. Then

resultant clusters are encoded as chromosomes and various operators are applied on the clusters: selection, crossover, mutation [11]. Here the Fitness function is the inverse of squared error value of the K Mean. Clusters with small squared value will have a large fitness value. The higher the fitness value, the higher will be the probability of clusters to be reproduced in next generation. Each entries of the chromosome denote the clusters to which the instance belongs. Each entry can have value 1 to K (K is the number of clusters). Steps of GA based clustering algorithm are: [9] **Input:** *S* (instance set), *K* (number of clusters), *n* (population size)

Output: clusters

Randomly create a *population* of *n chromosomes*;

Each instance belongs to a valid *K*-clusters of the data.

repeat

Calculate a fitness value of each chromosome

Regenerate a new generation of structures.

until

some termination condition is satisfied

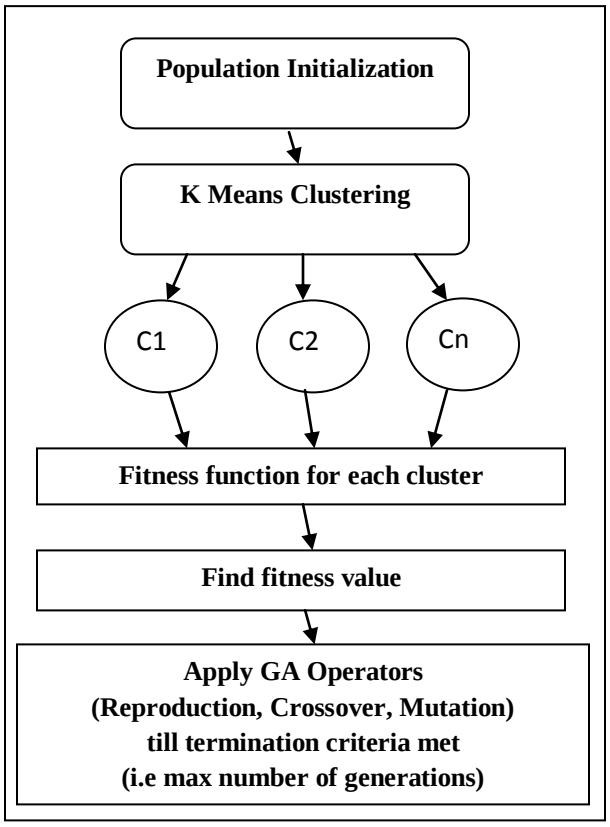

Figure 2: Genetic based clustering procedure

V. COMPARISON OF GA BASED WITH NON-GA BASED CLUSTERING ALGORITHM

Many non-GA-based clustering algorithms have been used such as K-means, Fuzzy-c-means in data mining. But non-GA-based clustering algorithms is not capable of efficiently forming natural groups for given large data set problems. This problem can be solved by employing an evolutionary approach i.e Genetic algorithm (GA) which is the best-known evolutionary techniques.[12]

Table 1: Comparison of GA – K Means – Fuzzy C Means Clustering

	GA Based Clustering	K-Means Based Clustering	Fuzzy C Means Clustering
Algorithms	Evolutionary based method	Partitioning based method	Partitioning based method with fuzzy clusters
Input	Number of clusters, population size, randomly chosen chromosomes, maximum number of iteration.	Number of clusters, dataset, randomly chosen clusters centroids.	Number of elements, Fuzzy clusters, Membership values w
Objective	Minimizing the sum of distances from each data point to its cluster centroid.	Minimizing sum of squared distance.	Minimizes intra-cluster variance as well.
Termination condition	When no. of iterations reached to maximum set value.	No changes in new cluster centroids.	When the coefficients' change between two iterations is no more than the given sensitivity threshold.
Search Approach	GA is based on global search approaches with implicit parallelism.	Final clustering may converge to local optima.	FCM is based on local search and it leads to local minimum.
Time Complexity	O(max*p*n*c*d)	O(n*c*d*i)	$O(ndc^2i)$

Where n=no. of data points, c= no. of clusters, d= dimension of data, max= maximum no. of iterations, p= population size.

VI. CONCLUSION AND FUTURE SCOPE

In this paper a review has been done on clustering technique with Genetic algorithm. It gives us a review on an efficient GA based clustering algorithm. For the creation of clusters, K Means algorithm is discussed and further GA is applied on clusters generated by K Means for its optimization.

There are other evolutionary techniques such as evolution strategies (ES), evolutionary programming (EP). In future, we can use these techniques in place of GA.

VII. REFERENCES

[1] Nikita Jain,Vishal Srivastava, "Data Mining techniques : A survey paper" , International Journal of Research in Engineering and Technology, pp. 116-119, 2013.
[2] M.S.B PhridviRaj, C.V. GuruRao, "Data Mining – Past present and future data streams," Elsevier, pp. 256-264, 2013.
[3] K.Kameshwaran, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining," International Journal of Computer Science and Information Technologies, pp.2272-2276, 2014.
[4] ShaloveAgarwal, ShashankYadav, Kanchan Singh, "K-means versus K-means ++ Clustering Technique", 2012 IEEE Second International Workshop on Education Technology and Computer Science.
[5] Lior Rokach, Oded Maimon. Data Mining and Knowledge Discovery Handbook, pp 321-352, 2005, Springer US.
[6] JaskaranjitKaur, GurpreetSingh,"Review of Error Rate and Computation Time of Clustering Algorithms on Social Networking Sites" International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 8, March 2015
[7] Goldberg, D. E, "Genetic Algorithms in Search, Optimization and Machine Learning", Reading, *MA: Addison-Wesley Publishing Co*, 1989.
[8] Byung-Jeong Lee, Byung-Ro Moon, Chi-Su Wu, "Optimization of multi-way clustering and retrieval using genetic algorithms in reusable class library".
[9] Hwei-jenlin*, fu-wen yang and yang-ta kao, "An efficient GA-based clustering technique", tamkang journal of science and engineering", vol. 8, no 2, pp. 113122 (2005).
[10] Manoj Kumar, Mohammad Husian, Naveen Upreti, Deepti Gupta, "Genetic Algorithm: Review and Application," International Journal of Information Technology and Knowledge Management, pp.451- 454, 2010.
[11] GunjanVerma, VineetaVerma, "Role and Application of Genetic Algorithm in Data Mining," International Journal of Computer Application, pp. 5-8, 2012.
[12] Rajashree dash and rasmita dash, "Comparative analysis of k-means and genetic algorithm based data clustering", international journal of advanced computer and mathematical sciences ISSN 2230-9624. Vol 3, issue 2, 2012, pp 257-265