Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.80 – 86

RESEARCH ARTICLE

A Comparative Analysis of Clustering Techniques using Genetic Algorithm

K.Bhuvaneswari[1], M.Anusha[2], Dr. J.G.R.Sathiaseelan[3]

Department of Computer Science, Bishop Heber college, Trichy 620017, India

E-mail address : bhuvanakrishnan.174@gmail.com [1], anusha260505@gmail.com [2], jgrsathiaseelan@gmail.com [3]

Abstract—- Clustering is one of the data mining techniques which could resolve most of the problems involved in data mining. The choice of clustering algorithm is based on the type of data that are used for a particular purpose and the relevant application. In order to improve the performance on unsupervised classification, evolutionary algorithm called genetic algorithm is applied on the data that could reveal the clustering issues like feature selection, cluster compactness, closeness, diversity production and so on. This paper presents a extensive survey on the performance of the clustering techniques based on the genetic algorithm.

Keywords— Data mining; Genetic algorithm; clustering techniques

I. INTRODUCTION

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

Traditionally, Clustering is one of the major techniques of data mining where similar characteristics of data are grouped. Therefore, the data on the same cluster shares similar patterns. The clustering overcomes classification by adapting the changes and the unknown features of the dataset. Now days, researchers concentrate on clustering owing to its properties such as scalability, ability to deal with different kinds of attributes, discovery of clusters with attribute shape, high dimensionality, ability to deal with noisy data, interpretability[1][2].

Various types of clustering techniques involved in data mining are center-based, density-based, and conceptual-based clustering techniques. Conceptual clustering is a machine learning paradigm for classification which distinguishes ordinary data by generating a concept of description for each generated classes. Most conceptual clustering methods are capable of generating hierarchical category structures. Conceptual clustering is closely related to formal concept analysis, decision tree learning, and mixture model learning. In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. The objects in these sparse areas are required to separate clusters and usually considered to be noise and border points. Hence a cluster will concise of all density-connected objects along with all objects that are within the objects range. Subsequently, center-based clustering represents clusters as a central vector, which might not necessarily be a member of the data set.

Even though, the process of clustering is identical to the classification, it differs from the choice of selecting data labels which is not known prior. The clustering algorithm can fetch most of the relevant objects at the same time irrelevant objects in the group. The process of identifying the unwanted data is called noise or outliners. A

good clustering method has to produce high quality clusters with maximum intra-cluster similarity with minimum inter-cluster similarity. The quality of a cluster could be improved with the help of the similarity measures and the cluster validity indexes. The optimality of the cluster could be decided by its viability of classifying the hidden patterns. The traditional clustering technique adequately falls into local optima which could be solved by incorporating genetic algorithm on the data sets [3].

Genetic algorithm guided by the Darwin's theory of evolution and natural genetics depicts the features like mutation, crossover, and the fitness evaluation, which could make it as a powerful tool to optimize the clustering techniques. Yet another feature of Genetic algorithm is population generation and maintenance of the generated population. The adapting suitable feature selection diversity of the population will be preserved [4].

The rest of the paper is organized as follows: Section II provides an overview of clustering with some of the algorithms. Section III contains the detail study of Genetic algorithm and genetic operators. Then section IV presents a massive study on clustering techniques with Genetic algorithm. Section V provides the performance analysis of the three algorithms. The conclusion of this paper is discussed in section V.

II. BACKGROUND

A. Clustering

The process of grouping objects into classes of similar object is called clustering or unsupervised classification. Clustering analysis helps to construct meaningful partitioning of a large set of objects, based on a 'divide and conquer' methodology. There are four different approaches to clustering such as partition clustering, grid-based clustering, hierarchal clustering, and density-based clustering. The partition clustering method groups N elements into K clusters and each cluster has at least one element. This method is suitable for small and medium size data sets. Grid-based clustering method defines the space of the database that is divided into a grid of separate cells. A single cell would be considered dense if it contains a large enough number of points which could initiate for the fast processing time. In this way, hierarchical clustering works by partitioning the data into the number of cluster in the sequential manner which has been projected as tree structured. Hence, the structure can be created with either a top down or bottom-up strategy. In top-down method, the database is considered at the beginning as a single big cluster and further divided into smaller clusters. The another method called bottom-up method which is propositional to the top-down method. Density based methods uses the concept of density to find arbitrarily shaped cluster. In contrast to density-based cluster, partitioning and hierarchical methods exhibits trouble in finding non-spherical clusters. The following are the reasons to choose clustering in various problems such as simplifications, pattern detection, data construction, unsupervised learning process and many more [5]. Some of the clustering techniques are described below.

B. K-means clustering

In this algorithm, the given data is divided into a pre-specified k clusters which alleviates the results by promoting at least one item in each cluster. This method avoids overlapping and non-hierarchical constructions. where one of them is the cluster centres. All data are processed and individual data points are assigned to the closest cluster center. This can be done using different measures of distance such as Euclidian distance. Based on the measures, a novel cluster centers are computed. These are the means of all cases assigned to each cluster center. The process is repeated until there are no remaining points to be assigned to different clusters over consecutive iterations, or until some other convergence criterion is reached. The goal of this processing is to assign points to clusters over successive iterations so that convergence to an optimal solution is reached.Convergence occurs when there are no other optimal solutions where points within a cluster are closer together, and cluster centers are farther apart. Figure 1 describes working of k- means clustering.

The weakness of this method could be the resultant solutions often falls into local optima. Hence global optimization is possible by adapting genetic algorithm which avoids local optimal solution, how many clusters we will have at the first [6].

C. CLARANS

Clustering Large Applications Based on Randomized Search algorithm was originated from two clustering algorithms such as PAM (Partitioning Around Medioids), and CLARA (Clustering Large Applications). CLARANS draw a sample with randomness in each step of the search. After the search the neighbours are shifted to current clusters. In case the algorithm identifies a good neighbor is found, CLARANS moves to the neighbor node and the process is started again, otherwise the current clustering produces a local optimum. The algorithm clusters all the data stored in main memory are considered to be the drawback of CLARANS clustering and this assumption may not be valid for large database [6].

D. Genetic Algorithm

Genetic algorithms are adaptive computational procedures modeled on the mechanics of natural genetic systems. The fundamental ideas of genetics where borrowed and used artificially to construct a better search algorithm. This algorithm is robust and requires minimum problem information for data processing. Normally, genetic algorithms tend to work better than traditional optimization algorithms because they are less likely to lost by local optima. This is because they do not make use of single-point transition rules to move from one single instance in the solution space to another. Instead, Genetic algorithm takes advantage of an entire set of solutions spread throughout the solution space, all of which are experimenting upon many potential optima. Genetic algorithm executed iteratively on a set of coded solutions, called population, with the three basic operators-selection, crossover, and mutation[7].

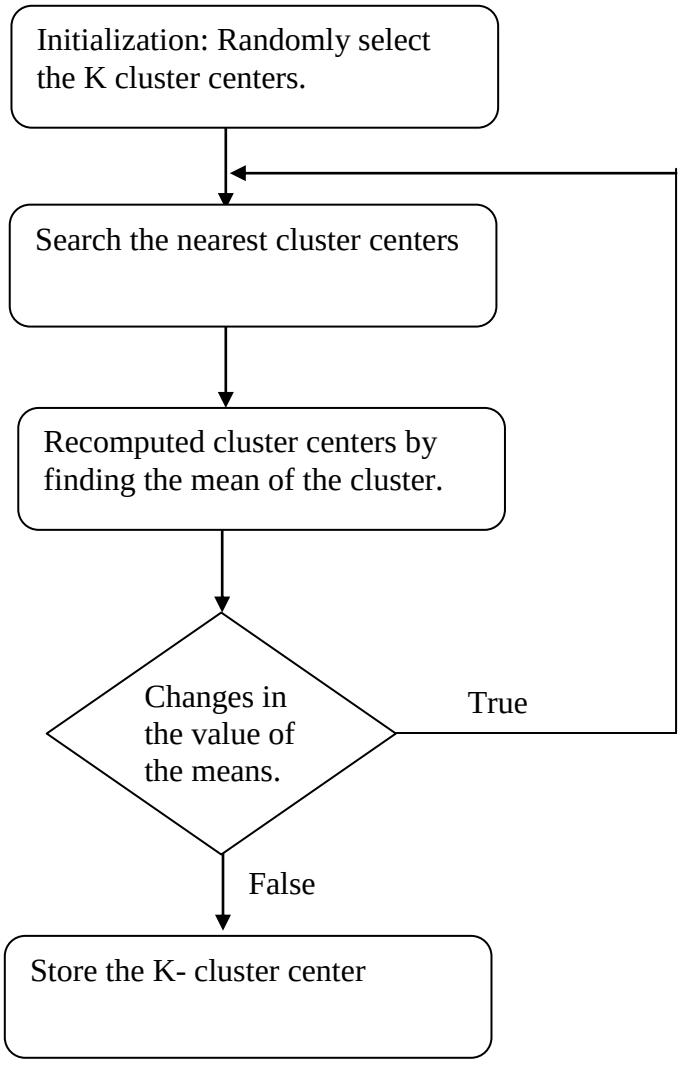

Fig.1. K-means algorithm

1)Selection:

GA should be able to produce the optimal or near-optimal solutions based on the selection scheme pressure. If the selection scheme pressure is too low, GA might take a long time for finding the optimal solution. In case when the selection scheme pressure is high, there is a chance for premature convergence of a suboptimal solution. There are many ways to select the individuals and two of the selection methods are discussed in this paper. They are roulette wheel selection and rank selection. In roulette wheel selection strategy, the chromosomes are selected based on its fitness value. In this process roulette wheel has many slots, and each slot, finds their chromosomes when the wheel randomly spun. These processes continue till all slots filled with

chromosomes. This method is not suitable for all problems, because the highly fit individuals may decrease the result. Another method named rank selection method selects chromosomes according to their achievements. The achievements are prioritized based on the total fitness of the chromosome. All the chromosomes are equally treated and the method is low in convergence rate compared to roulette wheel selection.

2)*Crossover*

Crossover is the recombination of two chromosomes which generates a new sub population from the original population by exchanging the information. There are various methods to crossover such as one point crossover, two point crossover, and n point crossover. In one point crossover approach, two parents are randomly selected and then swapped to create a new offspring. Figure 2 describes one point crossover operation.

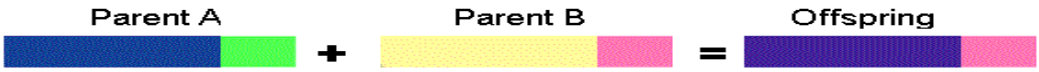

Fig. 2. One point crossover

Thereafter two point crossover randomly selects the middle parts of the parents chromosomes and then swapped to create a new offsprings. The procedure of selecting middle parents shows that the two point crossover is better than one point crossover.Fig.3 depicts two point crossover operation.

Fig. 3. Two point crossover

Another crossover called uniform crossover which specifies the mask bit value from the parent and transferred to the offspring. Figure 4 illustrates uniform crossover operation.

Fig. 4. Uniform crossover

The method arithmetic Crossover uses arithmetic operators for crossover rating. Most probable characters can be interchanged using logical operators and Fig. 5. explains the arithmetic crossover.

Fig. 5. Arithmetic crossover

3) *Mutation*

Mutation is another genetic operator which randomly changes the value of genes on a chromosome. If recombination is supposed to exploit the current solutions to find better ones, mutation is applied on the chromosome to help in exploration of the whole search space. Fig.6. describes the mutation process.

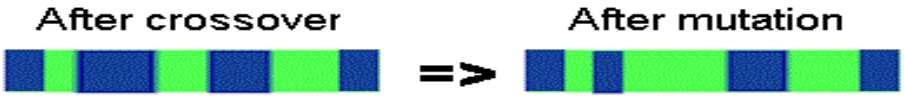

Fig. 6. Mutation process

4) *Termination*

The process of fitness computation, selection, crossover and mutation is continued for a fixed number of iterations or till the termination condition is achieved.

Pseudo code for genetic algorithm

Algorithm: Genetic Algorithm

1. Begin
2. Initialize population with random candidate solution
3. Evaluate each candidate
4. Repeat until termination condition satisfied DO
5. Select parents
6. Recombine pairs of parents
7. Mutate the resulting offspring
8. Select the individuals or next generation
9. End.

TABLE 1. GENETIC ALGORITHM PROCEDURE.

III. LITERATURE SURVEY

Agustin Blas et al.[8] described the performance of the grouping Genetic algorithm in clustering, started with proposed encoding, and different modification of crossover and mutation operation and also initiated the local search include with the island model for improve the performance of the problem. The real data sets like iris and wine were used and compared the results with the classical approaches such as DBSCAN and K-means, and obtaining the excellent results in proposed grouping based methodology the evolutionary approach such as Genetic algorithm. The performance of the algorithm was measured by using the different fitness function [8].

Tzung-Pei-Hong et al.[9] discussed the performance of the Genetic algorithm based attribute clustering process were improved based on the grouping Genetic algorithm. The chromosome representation, Genetic operations, and fitness function defined in grouping Genetic algorithm for solving the clustering problem. The result of grouping Genetic algorithm based clustering algorithm improved the convergence speed and fitness value of the clustering problem. In addition the algorithm can also deal with the problem of missing values. The other optimization algorithms are used to solve the problem in attribute grouping.

Daniel Gomes Ferrari et al. [10] proposed a new approach to characterize the clustering problems based on the similarity among objects and the method for combine internal indices for ranking algorithms based on the performance of the problem. The experimental results indicated the viability of meta learning systems for an unlabeled approach to the clustering algorithm selection problem. This technique presents the better result from the distance based set over the attribute based approach.

Edwin Aldana et al. [12] suggested a new approach for solved the searching space to find the optimal distribution of object in the clusters represents a hard combinatorial problem. Genetic algorithms increased the intra and inter clusters entropy simultaneously and yield the best possible combination of elements for a present number of clusters

Kunnuri Lahari et al. [13] enhanced reduce the local minima using evolutionary and population based methods like Genetic algorithm and teaching learning based optimization. The data sets iris and wine are used, and the experimental results are compared with the Genetic algorithm and teaching learning based optimization based clustering with k-means algorithm. The performance of the evolutionary based clustering method compared with some existing clustering method.

Rahila H.Sheikh et al.[14] proclaimed a brief study of Genetic algorithm based clustering. Rajashree Dash et al.[11] discussed on comparative analysis of K-means and Genetic algorithm based on clustering. Arun Prabha et al.[15] with respect to the idea were improved the cluster quality from K-means clustering using a Genetic algorithm. Large scale clustering problems in data mining also address by this method. The best results are achieved by using this method.

Anusha et al.[16] depicted an enhanced K-means Genetic algorithm for optimal clustering. The author overcomes the drawback of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the algorithm produced more than the 90% accuracy for real life dataset. The author also adopted a neighborhood learning strategy for optimizing multi objective problems. This algorithm used k means Genetic algorithm to find the compactness of the clusters. It is noted that the algorithm could produce minimum index value for the maximum datasets. However, there is a need for proper feature selection for better, more optimal solution [17].

IV. PERFORMANCE ANALYSIS

The results for the GGA(Grouping Genetic Algorithm) based clustering are compared with GA(Genetic Algorithm) based clustering. The values are shown in figure 7. From the figure it is identified that GA based clustering techniques sounds good for the real time datasets.

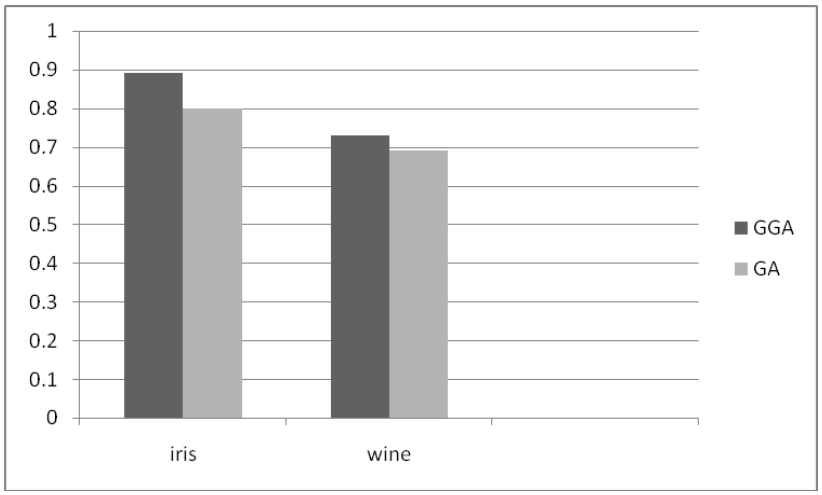

Fig. 7. Accuracy for real time dataset.

V. CONCLUSION

Clustering is one of the major task of data mining where the similarity patterns or similar objects a grouped by truncating the outliers. In this paper a survey has been taken on clustering technique with Genetic algorithm. It is inferred from the above study that an efficient clustering could be possible by incorporating the genetic algorithm into the cluster. Hence, by adapting clustering with genetic algorithm local optima could be avoided. The future aspect of this work could be application of medical data set on genetic algorithm with clustering to identify more relevant and compact clusters.

REFERENCES

[1] Nikita Jain,Vishal Srivastava, "Data Mining techniques : A survey paper" , International Journal of Research in Engineerning 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] Gunjan Verma, Vineeta Verma, "Role and Application of Genetic Algorithm in Data Mining," International Journal of Computer Application, pp. 5-8, 2012.

[5] Sharaf Ansari,Sailendra Chetlur, Srikanth Prabhu, N. Gopalakrishna Kini, Govardhan Hegde, Yusuf Hyder, "An Overview of Clustering Analysis Techniques used in Data Mining ," International Journal of Emerging Technology and Advanced Engineering, pp.284-286, 2013.

[6] Aastha Joshi, Rajneet Kaur, " A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, pp.55-57,2013.

[7] 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.

[8] L.E. Agust 1 n-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro-Calvo, J. Del Ser, J.A. Portilla-Figueras K. Elissa, "A new grouping genetic algorithm for clustering problems," Elsevier, pp.9695-9703, 2012.

[9] Honga Tzung-Pei, Chun-Hao Chenc, Feng-Shih Lin, "Using group genetic algorithm to improve performance of attribute clustering," Elsevier, pp.1-8, 2015.

[10] Danial Gomes Ferrari, Leandro Numes de Castro, " Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking combination methods," Elsevier, pp.181-194, 2015.

[11] Rajashree Dash and Rasmita Dash, "Comparative analysis of K-means and Genetic algorithm based data clustering," International Journal of Advanced Computer and Mathematical Sciences, pp.257-265, 2012.
[12] Edvin Aldana-Bobadhilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.
[13] Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Prediction based clustering using genetic algorithm and Learning Based Optimization Performance Analysis," Advances in Intelligent Systems and Computing," pp. 338, 2015.
[14] Rahila H. Sheikh, M. M.Raghuwanshi, Anil N. Jaiswal, "Genetic algorithm based clustering: A Survey," IEEE, pp.314-319, 2008.
[15] K.Arun Prabha, R.Saranya, "Refinement of K-means clustering using Genetic algorithm," Journal of Computer Application, pp. 256-261, 2011.
[16] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE, pp.580-584, 2014.
[17] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.

