International Journal of Engineering Studies.
ISSN 0975-6469 Volume 8, Number 1 (2016), pp. 37-45

http://www.ripublication.com

A Method for Clustering Large Data Deploying Multi-Objective Genetic Algorithm

Hoda Sadati[1] and Mohammad Sadati[2]

[1]Department of Engineering, faculty of Engineering, California State University, Los Angeles, America.

[2]Department of of Civil Engineering, faculty of Engineering, University of Southern California, Los Angeles, America.

Correspondence: Hoda Sadati. Tehran, Iran. Email: sadatii446@gmail.com

Abstract

Today, we are facing exponential growth of data. The data is used in medicine, social networks, and etc. Generally, we can say that the amount of data is increasing today. Therefore, analyzing and extracting useful information from such data is very challenging. Clustering plays an important role in managing large data and constitutes an important step in analyzing data. Because of their large size, one cannot easily categorize their parameters. Therefore, monitoring large data without knowing their nature is difficult. Deploying non-monitoring techniques like clustering can be very useful. These techniques utilize parameter learning for calculating teaching parameters. Clustering algorithms cab be used to label data to state similarity in a category. In this paper, we deployed multi-objective genetic algorithm to cluster large data. The algorithm functions better in clustering and in the time needed for calculation than other algorithms.

Keywords: Clustering, parameter learning, Clustering algorithms cab, genetic algorithm

1. INTRODUCTION

Large data, is frequent term and is associated with large data repositories (like data bases of organizations), for which common data processing methods do not work effectively [1]. As data generation grows around the world, complexities arise in analysis, selection, search, sharing, storage, transfer, visualization, privacy protection, and etc., the actualization of which is not possible with common methods [2].

Today, data have had great increase in terms of size and diversity, which makes their analysis relatively difficult. Clustering is a technique in which useful information and their hidden relation is assessed and discovered [3].
Clustering is one of primary data analysis techniques and is a tool for analyzing large data. Because of challenges associated with large data, different problems are faced when clustering them [2]. Since large data mean data in terabytes or petabytes, therefore, clustering them has large executing costs, analysis and study of algorithms effective in clustering and efficient in the time needed is an important issue [1].
Because of problems like processing speed, old clustering techniques are not applicable to large data. When working with large data, clustering algorithms should be able to perform the operation with a reasonable speed [4].
Clustering algorithms based on machine learning are able to identify clusters with arbitrary forms and deploying such methods for clustering network data had desirable results.
Methods and technologies deployed for working with large data and their effective analysis include parallel processing of databases, distributive file systems, data mining, etc. In this paper, utilizing multi-objective genetic algorithm, clustering operation of large data was performed and the effects of changing parameters of genetic algorithm on recommended algorithm function was assessed [5].
In the first part of this paper, large data are introduced and the definition of their clustering is stated. In the second part, reviews the literature and other related algorithms. The third part, introduces genetic algorithm. The fourth part introduces the recommended method and explains it thoroughly. The fifth part introduces the criteria for assessing the recommended algorithm and the assessment occurs in the sixth part. Parts seven and eight are dedicated to conclusions and references respectively.

2. LITERATURE REVIEW

Qing He et al (2014) discussed categorizations like partitioning, hierarchical, etc. and compared them. Their main focus was on the k-means algorithm. The k-means algorithm functions better than Mercer-Kernel based method and other algorithms [1].
Ganesh Krishnasamy et al (2014) were able to decrease k-means algorithm complexity by optimizing k-means which he did by deep learning [2].
Ishak Boushaki Saida et al (2014), deploying cuckoo search method, enhanced k-means method in a way that its inconsistencies were eliminated [3].
Xue-Feng Jiang et al (2014) promoted large data algorithms to decrease their operation time. Their algorithm performed based on particle optimization [4].
Khadija Musayeva et al (2014) introduced an algorithm to identify the number of clusters and optimizing their number which performed based on similarity matrix and could identify necessary clusters for categorizing clusters [5].
Hong Yu et al conducted studies regarding the number of clusters. Their model would determine the number of clusters based on the possibility of existence of one datum [6].
Wang Shuliang () analyzed large data clustering methods based on hierarchy network and categorized large data utilizing divide and conquer method [7].

J.V. Gautam (2015) analyzed scheduling algorithms using map reduce in hadoop and showed that deploying this framework causes enhancements in performance [8].
Gole (2015) assessed existing challenges using hadoop framework and introduced and algorithm for working with large data based on data mining [9].
Hoecker et al (2015) studied regarding large data categorization and authored papers introducing and analyzing different algorithms [10].

3. A REVIEW OF THE GENETIC ALGORITHM

This algorithm was introduced by John Holland in the Seventies. It functioned based on Darwin's Theory of evolution. There is an initial population consisting of chromosomes (strings of data) each composed of a series of genes. Genes (string elements) state the solution to the problem in code. Based on the conditions of the problems, a correspondence value is calculated for each chromosome [8]. Chromosomes which have higher correspondence value, survive and generate the next generation. This operation is known as survival of the fittest. Off springs of next generation are generated as a result of some operation performed on their parents. The operation consists of Selection, Crossover, Mutation, and Inversion [7].
To sum up, algorithms searching optimized results, are repeated processes. Here, too, first initial random values were chosen for chromosomes. Next, correspondence level was calculated for each of the chromosomes. The chromosomes were sorted based on descending correspondence. For higher level chromosomes, more options are considered for production of next generation [10]. Chosen chromosomes generate next generation and the process continues until specified conditions (acceptable answer to the problem or maximum repetitions) are met [9].

4. RECOMMENDED METHOD

In this paper, large data were clustered deploying genetic algorithm. Data for which clustering is done, were real data gathered from Tehran heart center. The data had parameters extracted from different people's ECG. They were associated with different people and consisted of nine characteristics for each person based on which one can determine heart health. The data were classified into two general classes, and for each class a specific label was determined. Therefore, there were two different labels namely "0" and "1" for different data in the dataset. If the label for a raw datum is zero, it shows that the heart disease didn't exist and the person was healthy. The number of data in the dataset was more than 195000 different data which were used as dataset to be clustered by the genetic algorithm.
In the recommended algorithm, normalization of data was first done and to increase speed, data were divided in five different equal parts, which are categorized in parallel. Utilizing paralleling when working with data caused the operation speed to be optimized. Then, results from the five parts were merged to produce final results. Work mechanism for clustering was the same for each of the five parts. In the following, the parts are looked into.

To cluster the data, we first had to determine the center for each part. To determine the centers, cost function FCM was deployed and the cost for each category was calculated and stored. Next, the data were merged and final results were obtained. To merge categories, first distance between centers were calculated and merging of categories was performed based on the calculated value, in a way that least distance between centers of data was considered as the distance between categories. In other words, we determined the distance between centers of data for the first and second class. This way, we could calculate the function for final costs for categorizing data. Next, most optimized classes or classes to which cluster center were closer, were determined and identified. It was obvious that the less the distance between clusters, the more optimized the algorithm did the clustering operation.

In the genetic algorithm deployed in this paper, the length of chromosome was equal to the number of data characteristics. In genetic algorithms, maximum number of repetitions is 200 times, and initial population is 500. The percentage of intersection was 0.8 and percentage of mutation was 0.02. The mutation operand performed by choosing a parent randomly and changing one gene. In intersection operand, single line and double line modes were use. Genetic algorithm was deployed to categorize data.

5. ASSESSMENT CRITERION FOR RECOMMENDED ALGORITHM

Different criteria can be utilized to assess recommended algorithm. One of the criterion deployed to analyze the algorithm, was the cluster internal distance in categorizing data. As mentioned beforehand, decrease in this distance indicated increase in accuracy of algorithm. Changing parameters like mutation or intersection would change the cluster internal distance and increase the accuracy of the recommended algorithm. Another criterion for analyzing the recommended method is the performance duration, which would change along with mutation or intersection parameters.

6. ANALYZING THE RECOMMENDED ALGORITHM

Figure 1 shows an example of performing the recommended algorithm in which decrease in cluster internal distance is observed. Figure 2, too, shows the results of data clustering. In this figure, red dots indicate data and blue dots indicate cluster center. Figures 3 and 4 show the change in mutation and intersection operands. In figure 5, the effect of change in intersection operand at the time of algorithm performance is shown and figure 6 shows the result of the change in mutation operand value on the time of algorithm performance.

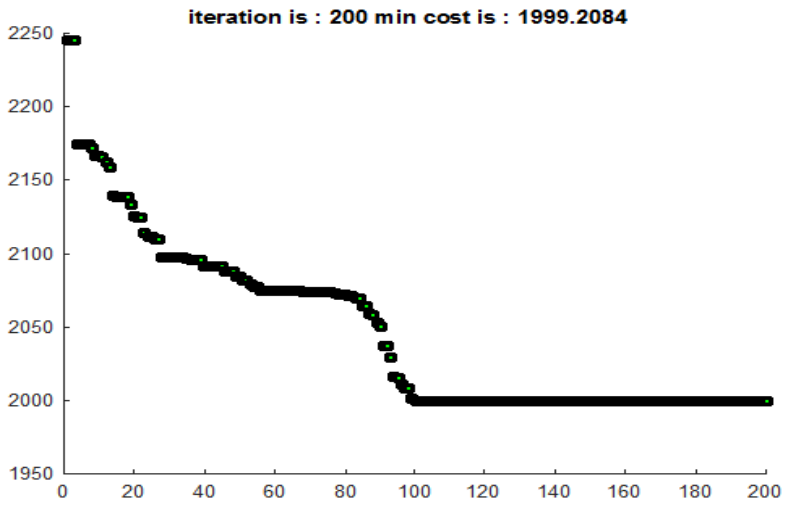

Figure 1: An example of performing recommended algorithm and calculation of cluster internal distance

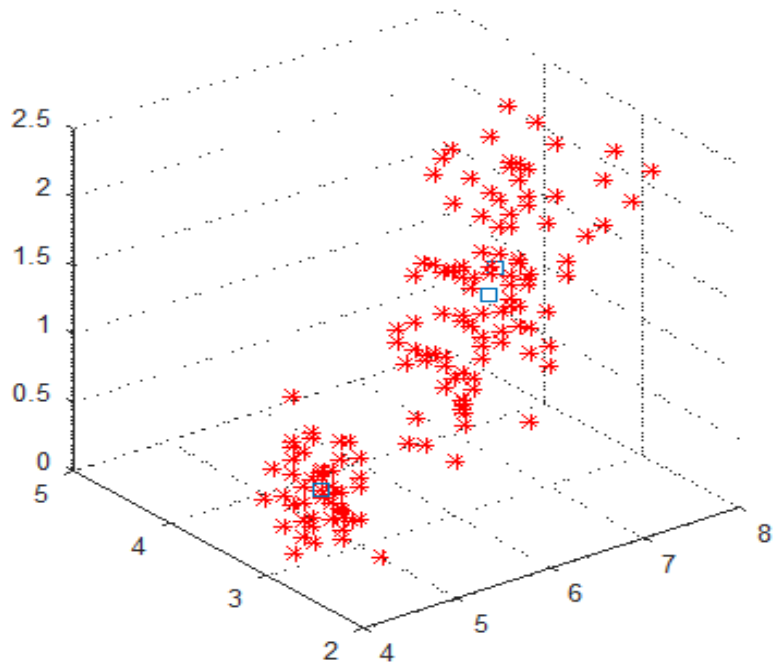

Figure 2: Showing results of data clustering

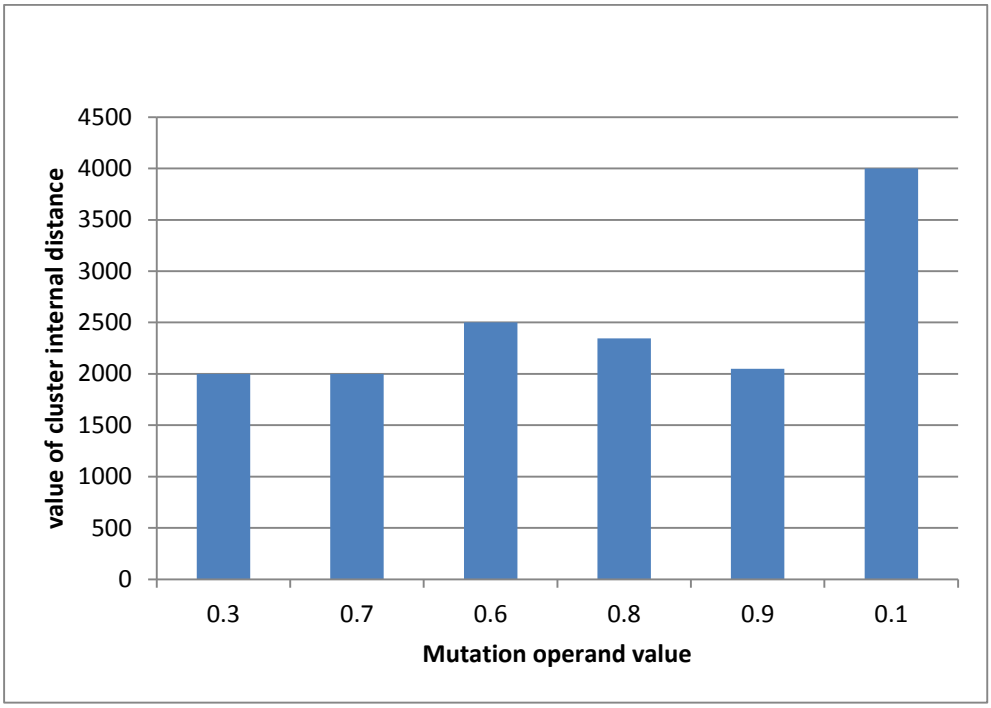

Figure 3: The results of mutation operand on cluster internal distance

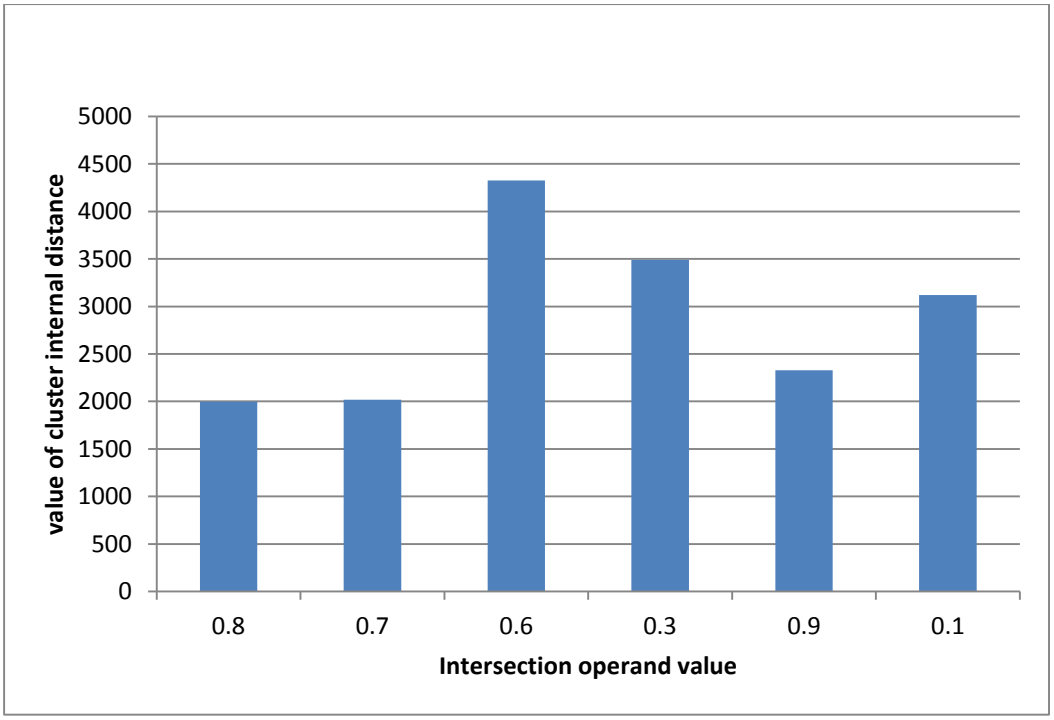

Figure 4: The results of intersection operand on cluster internal distance

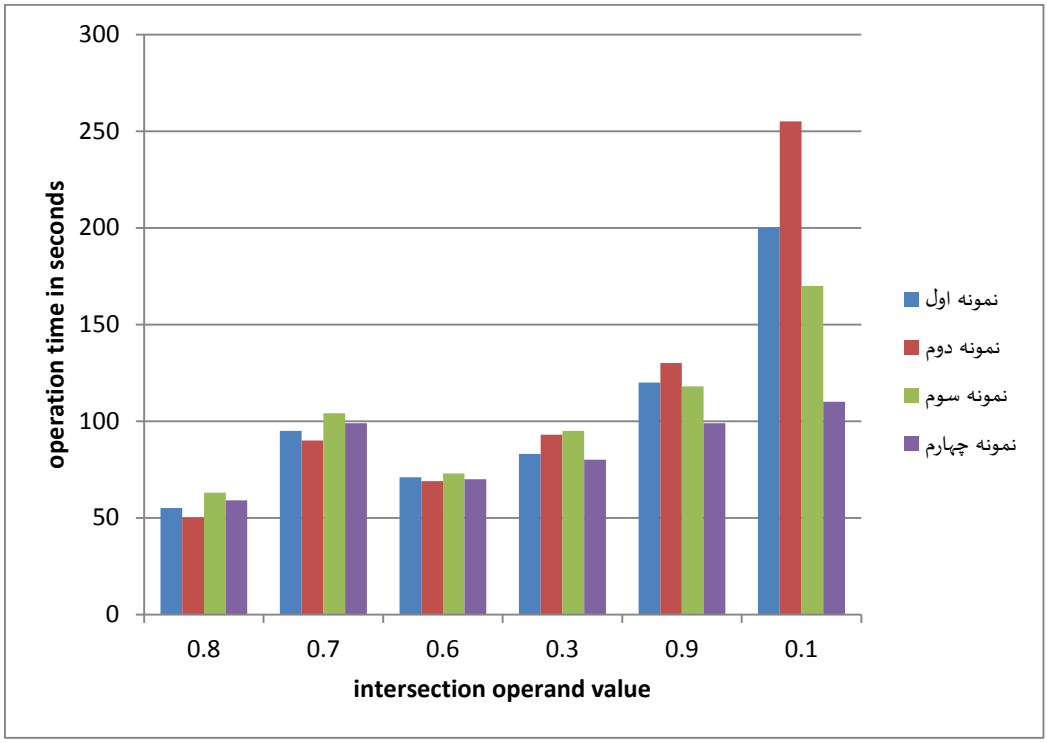

Figure 5: Operation speed after changing intersection operand value

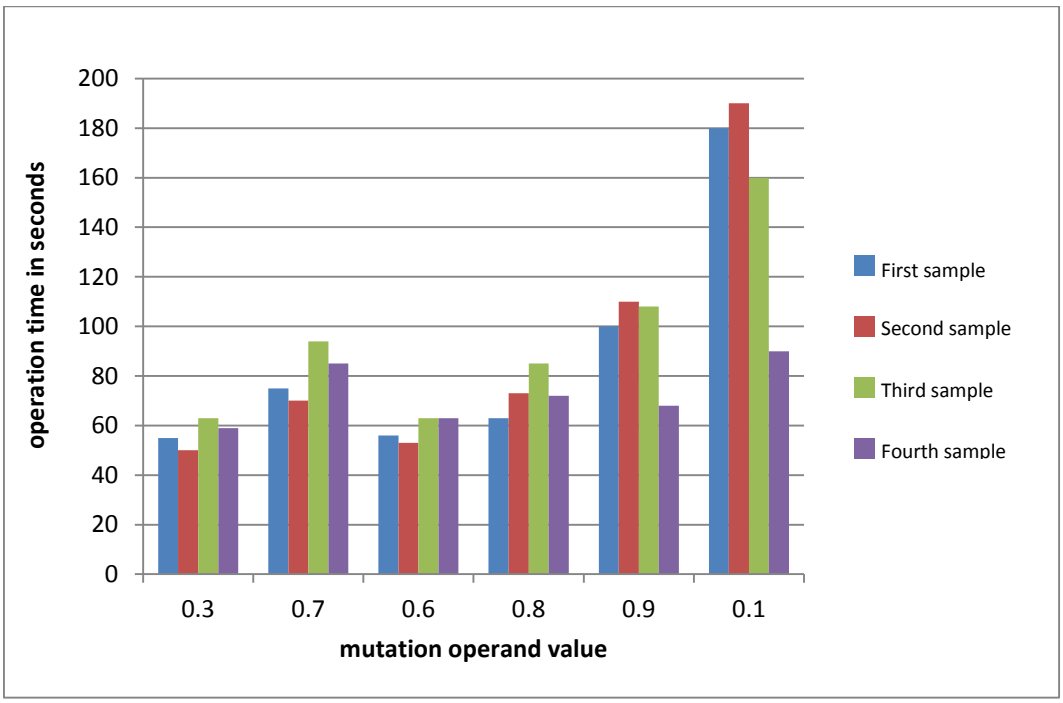

Figure 6 : Operation speed after changing mutation operand value

CONCLUSIONS

In this paper, genetic algorithm was deployed to cluster large data. Since genetic algorithm was an evolutionary algorithm, it could identify optimized points and do the clustering in a way that cluster internal distance would decrease

REFERENCES

[1] Qing He, Xin Jin, Changying Du, Fuzhen Zhuang, and Zhongzhi Shi Clustering in extreme learning machine feature space. Neurocomputing, 128:88{95, 2014.

[2] Ganesh Krishnasamy, Anand J Kulkarni, and Raveendran Paramesran. A hybrid approach for data clustering based on modified cohort intelligence and k-means. Expert Systems with Applications, 2014.

[3] Ishak Boushaki Saida, Kamel Nadjet, and Bendjeghaba Omar. A new algorithm for data clustering based on cuckoo search optimization. In Genetic and Evolutionary Computing, pages 55{64. Springer, 2014.

[4] Xue-Feng Jiang. Application of parallel annealing particle clustering algorithm in data mining. TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(3):2118{2126, 2014.

[5] Khadija Musayeva, Tristan Henderson, John BO Mitchell, and Lazaros Mavridis. Pfclust: an optimised implementation of a parameter-free clustering algorithm. Source code for biology and medicine, 9(1):5, 2014.

[6] Hong Yu, Zhanguo Liu, and Guoyin Wang. An automatic method to determine the number of clusters using decision-theoretic rough set. International Journal of Approximate Reasoning, 55(1):101{115, 2014.

[7] Shuliang WANG, Jinghua FAN, Meng FANG, and Hanning YUAN. Hgcudf: Hierarchical grid clustering using data _eld. Chinese Journal of Electronics, 23(1), 2014.

[8] J. V. Gautam, H. B. Prajapati, V. K. Dabhi and S. Chaudhary, "A survey on job scheduling algorithms in Big data processing," Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on, Coimbatore, 2015, pp. 1-11. doi: 10.1109/ICECCT.2015.7226035

clustering algorithm in data mining. TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(3):2118{2126, 2014.

[5] Khadija Musayeva, Tristan Henderson, John BO Mitchell, and Lazaros Mavridis. Pfclust: an optimised implementation of a parameter-free clustering algorithm. Source code for biology and medicine, 9(1):5, 2014.

[6] Hong Yu, Zhanguo Liu, and Guoyin Wang. An automatic method to determine the number of clusters using decision-theoretic rough set. International Journal of Approximate Reasoning, 55(1):101{115, 2014.

[7] Shuliang WANG, Jinghua FAN, Meng FANG, and Hanning YUAN. Hgcudf: Hierarchical grid clustering using data _eld. Chinese Journal of Electronics, 23(1), 2014.

[8] J. V. Gautam, H. B. Prajapati, V. K. Dabhi and S. Chaudhary, "A survey on job scheduling algorithms in Big data processing," Electrical, Computer and

Communication Technologies (ICECCT), 2015 IEEE International Conference on, Coimbatore, 2015, pp. 1-11. doi: 10.1109/ICECCT.2015.7226035

[9] S. Gole and B. Tidke, "A survey of big data in social media using data mining techniques," Advanced Computing and Communication Systems, 2015 International Conference on, Coimbatore, 2015, pp. 1-6. doi: 10.1109/ICACCS.2015.7324059

[10] M. Hoecker, K. L. Polsterer, S. D. Kügler and V. Heuveline, "Clustering of Complex Data-Sets Using Fractal Similarity Measures and Uncertainties," Computational Science and Engineering (CSE), 2015 IEEE 18th International Conference on, Porto, 2015, pp. 82-91. doi: 10.1109/CSE.2015.35