RESEARCHERID
THOMSON REUTERS

[ICESTM-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

DEALING BIG DATA USING FUZZY C-MEANS (FCM) CLUSTERING AND OPTIMIZING WITH GRAVITATIONAL SEARCH ALGORITHM (GSA)

Mr R Venkat[1], PhD Scholar & Dr K Satyanarayan Reddy[2], Professor

[1&2]Cambridge Institute of Technology (CITECH), Bangalore, Visvesvaraya Technological University, Belgaum, Karnataka State, India.

ABSTRACT

Any data in the real world can be organized properly by Clustering the data using some clustering techniques and in these clustering techniques Fuzzy C-Means (FCM) is a very recent and better technique that can mold the data with good logic and in a highly accurate manner. FCM is same as K-Means clustering technique but FCM is developed with some Fuzzy means. FCM is associated with some constraints as FCM is more responsive as on the order how the clusters were arranged in the beginning of applying the technique. Even though it has its own defects when dealing with large data, it cannot get the finest solution as the data need some optimization? So to get some better accuracy, in this paper we go with an improved FCM Clustering Algorithm technique where for the accurate solution we involve Gravitational Search Algorithm (GSA), which is based on Law of gravity and notion of masses. GSA can be used to improve the limitations in order to gain a well-defined system performance, GSA can efficiently deal with many propositions of large data. Here we are going to develop a Map-Reduce mechanism to effectively deal with large data. The expected result can be achieved by choosing the better finest candidates and grouping them in the reduction mode to get the finest solution. Evaluation of practical results will be optimizing when compared to other existing similar techniques.

I. INTRODUCTION

Clustering is an activity in which we align data items where similar data items are assembled in to one set known as clusters. Moreover clustering is a mechanism which gives some unproven results, here are all data items after clustering belonging to a single cluster can be considered as an identification issue [5]. Any clustering mechanism can give proper results only if the datasets are organized properly such that all similar datasets can be grouped into a single category and also the datasets in one category should have similar properties with each other and other datasets in other category will vary [1]. Here clustering mechanism used is Fuzzy C-means (FCM) clustering which is a very latest technique, FCM assigns membership to every data point in every cluster centroid by the distance between the centroid and the data point. If data is high to centroid, then membership is high. A data point within a cluster can belong to any number of clusters; each data point will have a probability or membership of belonging to each cluster. When the probabilities or memberships of a data point are added then the result should be one. Some complications associated with FCM are heavy responsiveness towards early conditions of Centroid, less possibility to obtain better optimum results globally when we are dealing with large data [3].

For Optimization we are involving Gravitational Search Algorithm (GSA), GSA is based on law of gravity and notion of mass interactions. When applying GSA, gives better results by identifying optimum number of clusters and curtails the fitness function and also we are involving Map-Reduce mechanism for a high level methodology for Large data.

II. OPTIMIZED FCM ALGORITHM

As there many complications associated with normal FCM technique, now we go with some optimized FCM by involving GSA optimization which can deal large data. In a normal GSA, the particles velocity is with only one value, to obtain all values we need to go in all propositions of data, for that we need assign a velocity for every data item. These practical works will exhibit a high level précised results. A Conventional FCM can fit only for normal optimizing process when dealing with large data, FCM is highly responsive to the early arrangement of Centroid

ISSN 2348 – 8034
Impact Factor- 5.070

which affect the convergence. After this GSA can optimize the process globally and converge better. This is a good methodology of merging to techniques to enhance the values. This is a high level technique that uses cluster centre as items and utilize GSA and FCM for results extraction. Fig. 1 shows the FCM mechanism. Here each item represents a candidate solution for the case and their location is corresponding to the centroids and gives the vector for all the centriods. Evaluation of fitness function is to minimize and obtain the best result.

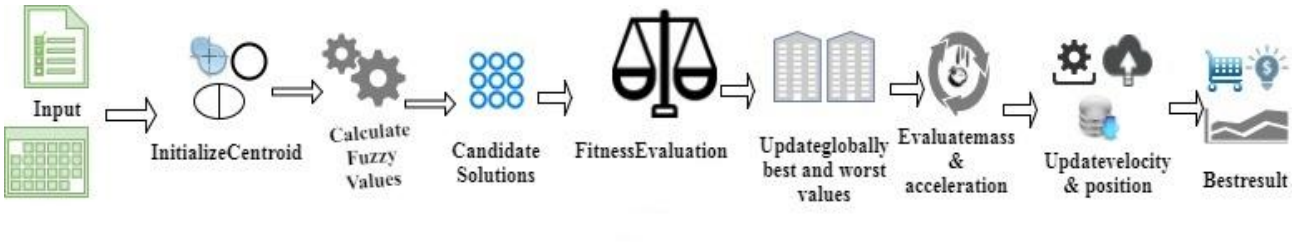

Fig.1 Fuzzy C-Means Clustering (FCM) [17]

Fig. 2 shows the GSA mechanism, here we are merging FCM and GSA, GSA is for searching globally and FCM for locally, and the main thing is to involve FCM when dealing with GSA. Here the proposed system begin with identifying cluster centers by involving GSA with FCM and set of data items converge at a position, corresponding to the constraints for which the difference of fitness values for identifying items goes to threshold, for this we go with fitness variance.

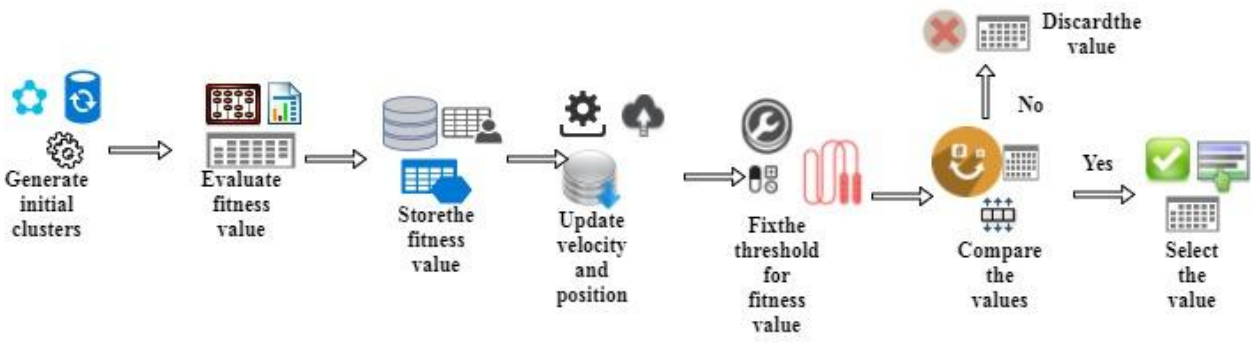

Fig. 2 Gravitational Search Algorithm (GSA) [17]

III. MAP-REDUCING MODEL

Here we are using Map-Reducing Model for parallelizing and enhancing the clustering speed and efficiency. First we need to divide the large data in to partitions; these partitions can same or different in size, later every partition is mapped, this mapping gives some centroids. Centroids will be in the best possible types specifically. The mapping output is combined in the grouping level, which make them into a reduction process. Here in the reducing process, some centroids are termed as candidate optimal centroids. For every round of reducing, standard deviation of centroid result is evaluated. The process is completed when the preferred threshold is gained otherwise the mapping will be done until we get the expected result as shown in Fig. 3. The main thing in this process is selecting keys optimally; the effective thing in this process depends on how the data items are distributed and on how the keys are selected. So in this proposal, key is the file names in the first round, the next round is for better and fast similarities, we need to use the extracted centroids. The data partitions which are closer to the evaluated centroids will be subjected to the similar reducing process. Next for the combining process , we invoved FCM with GSA which gains the evaluations in a subjective way. So for every round, computing values run in analogous and data is combined.

ISSN 2348 – 8034
Impact Factor- 5.070

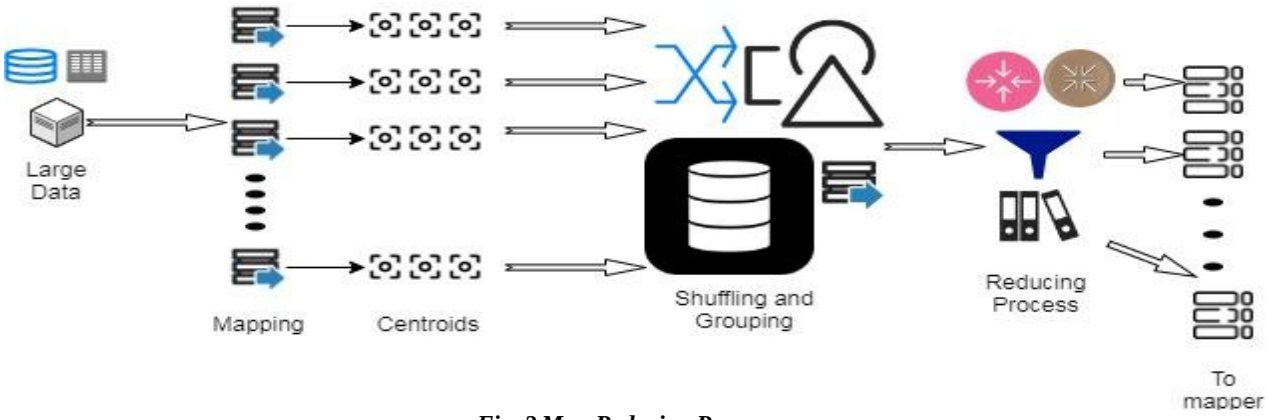

Fig. 3 Map-Reducing Process

IV. CONCLUSION

Here in this proposal we are developing an optimized FCM mechanism which can conquer the defects in normal FCM like over responsiveness to early arrangement of clusters and also less convergence. For this we are merging FCM with GSA optimizing process and go with every proposition of data to gain précised results. Even we are using Map-Reduction process to deal with large data. Using these combinations give better results when dealing with large data. Even we can use these mechanisms for dealing with social networks data and medical data.

REFERENCES

[1] *A. K. Jain, M. N. Anil, P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), 1999.*
[2] *J. C. Bezdek, R. Ehrlich, W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, 10(2), 1984, 191-203.*
[3] *Y. Xianfeng, L. Pengfei, "Tailoring fuzzy C-means clustering algorithm for big data using random sampling and particle swarm optimization," International Journal of Database Theory and Application 8.3, 2015, 191-202.*
[4] *E. Mehdizadeh, S. Sadi-Nezhad, R. Tavakkoli-Moghaddam. "Optimization of fuzzy clustering criteria by a hybrid PSO and fuzzy c- means clustering algorithm," Iranian Journal of Fuzzy Systems 5.3, 2008, 1-14.*
[5] *A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters 31, no. 8, 2010, 651-666.*
[6] *E. H. Ruspini, "A new approach to clustering," Information and control 15.1, 1969, 22-32.*
[7] *G. H. Liang, T.Y. Chou, T.C. Han, "Cluster analysis based on fuzzy equivalence relation," European Journal of Operational Research, 166(1), 2005, 160-171.*
[8] *M. S. Yang, H.M. Shih, "Cluster analysis based on fuzzy relations," Fuzzy Sets and Systems, 120(2), 2001, 197-212.*
[9] *Z. Wu, R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 15(11), 1993, 1101-1113.*
[10] *J. C. Bezdek, J.D. Harris, "Fuzzy partitions and relations; an axiomatic basis for clustering," Fuzzy sets and systems, 1978.*
[11] *A. O. Esogbue, "Optimal clustering of fuzzy data via fuzzy dynamic programming," Fuzzy Sets and Systems, 1986.*
[12] *D. Graves, W. Pedrycz, "Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study," Fuzzy sets and systems, 161(4), 2010, 522-543.*
[13] *C. Ordonez, E. Omiecinski, "FREM: fast and robust EM clustering for large data sets," ACM, 2002.*

ISSN 2348 – 8034
Impact Factor- 5.070

[14]T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: an efficient data clustering method for very large databases," In ACM Sigmod Record, ACM, 1996.

[15]J. Cervantes, F. García-Lamont, A. López, L. Rodriguez, J. S. Ruiz Castilla, A. Trueba, "PSO-Based Method for SVM Classification on Skewed Data-Sets," In Advanced Intelligent Computing Theories and Applications, Springer International Publishing, 2015.

[16]Amir Khoshkbarchi, Ali Kamali, Mehdi Amjadi, Maryam Amir Haeri, "A Modified Hybrid Fuzzy Clustering Method for Big Data," In 2016 8th International Symposium on Telecommunications (IST'2016), 20016.

[17]R Venkat, Pamidi Srinivasulu, " Clustering of data using fuzzy C-means (FCM) algorithm with aid of gravitational search optimization," 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017.