Clustering of Huge Data with Fuzzy C-Means and applying Gravitational Search Algorithm for Optimization

R Venkat, K Satyanarayan Reddy

Abstract: ***There is a lot of bulk data which can be efficiently structured using some Clustering mechanism, among these mechanisms Fuzzy C-Means (FCM) Clustering technique is very new and can handle this bulk data logically and in a well precise mode. FCM is a better technique when compared to K-Means as FCM is designed with Fuzzy Concerns. But clustering only cannot give precise outcome, that's the reason we are involving an Optimization technique for tuning the results and Gravitational Search Algorithm (GSA) Optimization can makes the outcome more precise. GSA is concerned with gravity principles. GSA tailors the defects and transitions into a well structure system and finally FCM will be optimized using GSA. This System is developed with Map-Reduced method. Here in this paper, a discussion is being presented with different existing techniques that were previously used to structure the data and it is discussed how FCM with GSA is better technique when compared to those techniques and some sample Preprocessing Patterns and k-means clustering results are obtained as a first step of research.***

Keywords: ***Fuzzy C-Means Clustering, Gravitational Search Algorithm, Huge Data, K-Means and Preprocessing.***

I. INTRODUCTION

In the previous research studies [1], it is stated that using Fuzzy sets is a new way of representing reduced data and this minimizes the issues in clustering data and provides better understanding of actual data's format. Here the idea of clustering methods is to discrete things in a manner such that all things which are closer to each other are taken in to one set [2]. Any Clustering technique can be precise if it can give the least average distance between the data points within the same group and also if it gives the higher average distance between the data points of different groups [3].

In Fuzzy based clustering, a data point does not belong to a single group, the data point can fit in multiple groups and every instance (data point) will be assigned a probability value with each group, such that adding probability values of an instance gives one as it is a single data point [4]. But Fuzzy Clustering has some issues such as heavy responsiveness to previous stage i.e. early formation of cluster centroids and there is assurance to get accurate results when dealing with huge data [5]. There is substantial growth of web services, many internet-based sources like huge data and most effective data tools were developed. Lot of different data in many formats is created by public, regarding public and with many other sources associated with them. Various groups suggests the advantages of clustering data from different social networks where lot of public interact and post their data. The content is from many internet sources available for many users [6]. The technology growth made acquiring data simple; this made people to keep a collection of lots of data. Even using digital data they are retrieving interesting patterns. Using data mining techniques data can be manipulated and an entity's nature can be known time to time [7]. Considering any data is moulded with some Clustering mechanism and FCM is a new and proper mechanism that moulds data with proper measure and precise mechanism. FCM is concerned with Fuzzy concerns, highly reactive to starting arrangement of Groups. And it poses some demerits even, for obtaining proper results there should be some optimization. For Gravitational Search Algorithm (GSA) is involved, GSA works on the basis of Gravitational Principles and insight of Masses. This improves the demerits of GSA. Map-Reducing deals the process better by getting the better solution [16].

II. METHODOLOGY

Applying mining on Community data and on any falsified data, which is some wide variety of data with business processing, user interests [8]. Large-scale data processing has recently emerged as an increasingly important challenge, as it has become a salient operation for a wide variety of applications. Such data-intensive applications include web-scale data processing (e.g., web search and indexing) as well as solutions to a multitude of scientific problems by processing vast amounts of continually changing data [9]. Information application services permit normal and predictable huge data collection, which are useful for preparing some mining principles.

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.
* Correspondence Author

R Venkat*, Assistant Professor , DVR & Dr HS MIC College of Technology, Kanchikacherla, Dist. Krishna.

K Satyanarayan Reddy, Head of the Department, Information Science & Engineering, Cambridge Institute of Technology, Bangalore, Karnataka State, India.

In the primary view, processing decisions can lead to positive sense; Classification cannot give proper results by personal interests [10]. Big data is very complex content which is typical to go with customary methods [11]. Clustering is a method in which large complex data is partitioned to groups as the instances in one group will be very close. Any such data can be processed using Clustering and we have such applications in a huge number [12].

Here analyzing Visuals makes users to get more knowledge, but these techniques are not up to the mark [13]. Analyzing data is a challenging task in current trends; Applications of investigation are discovering weather outlines globally, commercial improvements and social issues. Cloud technology involving software content like Map-Reduce outline gives a comeback to the issue with distribution simulations around huge nodes. In many cases, input is physically distributed around central source, and moving entire content to one source is very costly. The mechanisms whatever discussed are to implement the task with only one centroid and can have a poor response when applied. [11] Here many probable methods of implementing such tasks are examined, later statistics with data transformation are proposed and these can utilized for defining the time plans for task sequences. Analyzing huge data sets is crucial for better results. Here the main task is determining the nearest cluster which satisfies the constraints and symbolizes the examined data. [14] Have proposed Approaches for nearest neighbor and comparing with existing results. For matching features, there are two better techniques they are k-d forest and k-means search tree with priority. Here Cloud Computing a capable technique for such data with high processing ability with high storage, scalable issues, and resource reusable capabilities with fewer budgets is suitable for high data. On the opposite side, one cloud can hold a variety of services, cross and internet based services are heavily time-consuming for such applications. It encounters effectiveness of service content growth in the internet. Here there bandwidth problems for web sources, within a scalable application it is a tough task to manage because of cross cloud combination as the time cost is high for these applications and this will be a target to process such data.

The various problems that exist in the recent works are,

- ➢ Primarily, the outcome from these techniques varies among many and goes with early conditions. And the variation increases when results increase as the data has many dimensions.
- ➢ Second, the algorithms used are time-consuming for large datasets.

[2] In Clustering all instances are formed to groups where all similar instances are considered in to one group, and these are not as accurate as they are considered as a reference. [3] Among different clustering techniques only techniques that can mould data properly are trustworthy like FCM. This assigns probability value for every instance, when probability value is high as the data towards the instance is far. As discussed FCM works with probability measure and characterized with some demerits i.e. obtaining precise results is difficult [5]. GSA can fulfill the difficulties in achieving precise results. FCM is rectified with GSA; Fig.1 is FCM technique which valuates Fitness to get precise results [16].

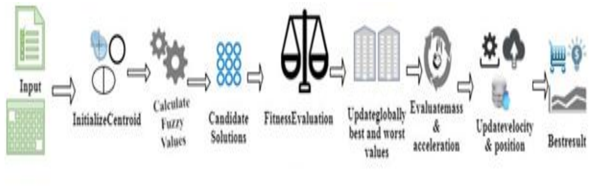

Fig.1. Fuzzy C-Means Technique [16]

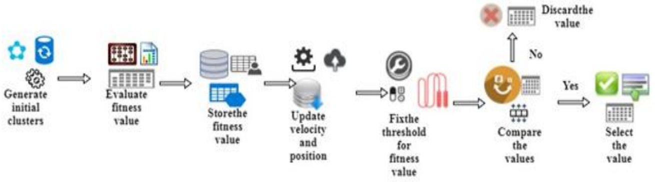

Fig. 2. Gravitational Search Algorithm [16]

Fig.2 GSA process, FCM with GSA finds Cluster Centroids and fitness variance is required. Map-Reduction process is for Parallelize and improves the FCM technique and makes it effective. Initially it partitions the Huge Data into similar or dissimilar parts and maps them results as Cluster Centroids, which are finest. Outcomes obtained are integrated with the groups and results as Reduction, where the Cluster Centroids are the best nominees. For each turn standard deviation is evaluated until the precise result is obtained in Fig. 3. Keys Selection plays vital role in the process which depends on the way how instances are dispersed. Partitions nearer to the centroids obtained are given to reduction. Enhanced FCM involved with GSA obtains improved outcomes [16].

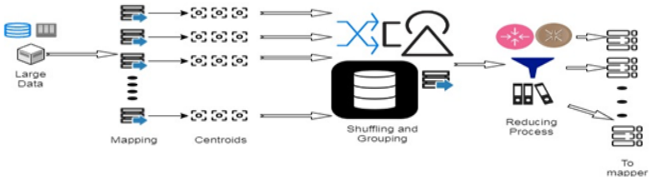

Fig. 3. Map-Reducing Process [16]

[14] Proposes Approach for estimated closest neighbor for matching applying k-means, calculating and comparing these with similar existing approaches for getting good features. The drawbacks in k-means is Hierarchical nature, it produces many divisions and here each instance need to belong to a single group but in [16] FCM an instance can belong to any number of groups as discussed. [15] Proposed Confidential based combined cloud methodology, k-means was used as data retrieving tool to gather essential information which could protect cloud privacy, [18] the main disadvantage of k-means is it fails in determining number of clusters before handling and other disadvantages are all data forced in to a cluster, Clustering models can depend on starting locations of clustering centers, needs k clusters to be determined initially and detects only hyper spherical clusters that are well separated and Sensitive to noise & outlier points. But [16] FCM has Objective function to select optimum number of clusters as this gives probability value. [17] States that for tailoring FCM to overcome limitations,

GSA which is motivated with Gravity principles is applied. GSA with FCM is able to identify better clusters.

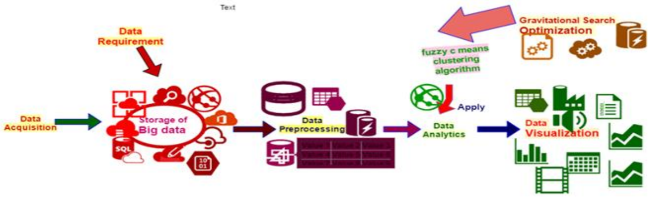

Fig. 4. Proposed System Architecture

III. PREPROCESSING AND K-MEANS RESULTS

As an initial step of research, a dataset in UCI data repository is preprocessed, Data preprocessing is a process which is used to change the data from raw format to proper usable format and used to clean the data by removing noise or outliers from data such that the data will be ready to apply any mechanism. For this task, a data set from 'User Identification From Walking Activity' is chosen, these datasets are having the data regarding users walking activity and Data Preprocessing is done for this particular data using Weka data mining tool and some sample preprocessed patterns are gathered. And the data will be preprocessed, means the noise will be removed from the data and the data will be ready for applying the mechanism.

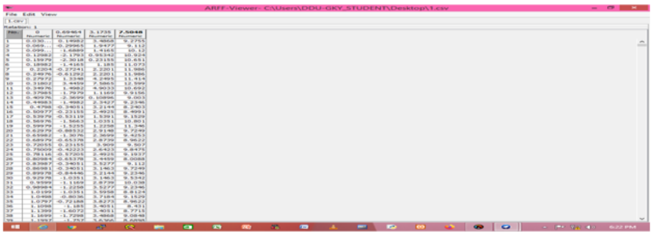

Fig. 5.Dataset loaded in Weka tool

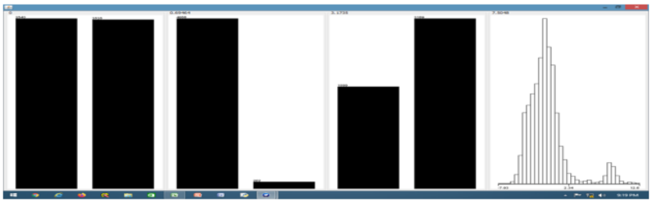

Fig. 6. Preprocessed Patterns

The above patterns are the Preprocessing Patterns obtained using Weka tool. The results below are k-means results whatever obtained by applying simple k-means in Weka tool for the dataset after preprocessing.

Running of k-means

Instances obtained: 5068

Attributes obtained: 4

0
0.69464
3.1735
7.5048

Number of iterations: 2

Within cluster sum of squared errors: 2202.8320961769487

Missing values globally replaced with mean/mode

Table- I: Cluster Centroids obtained by k-means

	Cluster#			
Cluster Centroids	*Attribute*	*Full Data*	*0*	*1*
		-5068	-2543	-2525
	0	'(-inf-77.115319]'	'(-inf-77.115319]'	'(77.115319-inf)'
	0.69464	'(-inf-0.8037]'	'(-inf-0.8037]'	'(-inf-0.8037]'
	3.1735	'(7.893-inf)'	'(7.893-inf)'	'(7.893-inf)'
	7.5048	-0.8457	-0.9471	-0.7436

Table- II: Model and evaluation on training set

Evaluation	Clustered Instances	
	0	*1*
	2543 (50%)	2525 (50%)

The total Instances are 5068 and the attributes are 4, Table-I represents the Cluster Centroids obtained by k-means and Table-II is regarding Evaluation on training set with Clustered Instances. In the next step, discussed Mechanism i.e. FCM with GSA is applied on the preprocessed data and some better clusters will be obtained which will be better than existing similar results.

IV. CONCLUSION

A System is being designed in order to cluster the big data using advanced Fuzzy c means Algorithm with the aid of Gravitational Search Optimization. The result obtained is related to the number of cluster heads and the clustering efficiency using GSA optimization process. Finally, the applicability of the proposed algorithm will be checked for data clustering using different data available in the UCI machine learning repository. With these results investigation is done for increasingly large volume of datasets in the Cloud and a plan is done to investigate the Scalability of Privacy Preservation in Big Data Applications with Cloud Service Access. As an initial step some Preprocessed Patterns and k-means results are obtained. In the next step, the discussed mechanism (FCM with GSA) will be applied on the preprocessed data and some better results are obtained. The proposed work will be compared with existing clustering algorithms and optimization algorithms in terms of clustering accuracy with respect to Sensitivity, Specificity and cluster size etc.

REFERENCES

1. E. H. Ruspini, "A new approach to clustering," Information and control 15.1, 1969, 22-32.
2. A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters 31, no. 8, 2010, 651-666.
3. A. K. Jain, M. N. Anil, P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), 1999.
4. J. C. Bezdek, R. Ehrlich, W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, 10(2), 1984, 191-203.
5. 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.
6. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Zomaya, S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE Transactions on Emerging Topics in Computing,Vol.2 ,No.3,2014.

7. Oren Shmiel, TomerShmiel, Yaron Dagan, and Mina Teicher, "Processing of Multichannel Recordings for Data-Mining Algorithms",IEEE Transactions On Biomedical Engineering, Vol. 54, No. 3,2007.
8. Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo and Chengqi Zhang, "Combined Mining: Discovering Informative Knowledge in Complex Data", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 41, No. 3, 2011.
9. R. Jain,P. Sarkar and D. Subhraveti ,"GPFS-SNC: An enterprise cluster file system for Big Data", IBM J. RES. & DEV, VOL. 57 NO. 3/4, 2013.
10. Sara Hajian and Josep Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining",IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 7,2013.
11. Lizhe Wang,Hui Zhong, Rajiv Ranjan, Albert Zomaya And Peng Liu, "Estimating the Statistical Characteristics of Remote Sensing Big Data in the Wavelet Transform Domain", IEEE Transactions on Emerging Topics in Computing, Vol.2 ,No.3,2014.
12. Timothy C. Havens, James C. Bezdek, Christopher Leckie, Lawrence O. Hall and Marimuthu Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data", IEEE Transactions On Fuzzy Systems, Vol. 20, No. 6,2012.
13. Jaegul Choo and Haesun Park, "Customizing Computational Methods for Visual Analytics with Big Data", IEEE Computer Society,2013.
14. Marius Muja and David G. Lowe, "Scalable Nearest Neighbor Algorithms for High Dimensional Data",IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 36, No. 11, 2014.
15. Wanchun Dou, Xuyun Zhang, Jianxun Liu, and Jinjun Chen, "HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications", IEEE Transactions On Parallel And Distributed Systems, Vol. 26, No. 2, 2015.
16. R Venkat, Dr K Satyanarayan Reddy, "Dealing Big Data using Fuzzy C-Means (FCM) Clustering and Optimizing with Gravitational Search Algorithm (GSA)", 3rd International Conference on Trends in Electronics and Informatics, IEEE, pg no 465-467, 2019.
17. Ananthi Sheshasaayee ; D. Sridevi, "Fuzzy C-means algorithm with gravitational search algorithm in spatial data mining", 2016 International Conference on Inventive Computation Technologies (ICICT).
18. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

AUTHORS PROFILE

R Venkat, done B-Tech(CSE) in MIST, Sathupally, Khammam Dist., AP, India from 2004-08 and M-tech(CSE) in K L University, Vaddeswaram, Guntur Dist., AP, India from 2009-11. He is currently pursuing Part time PhD in Viswesvarayya Technological University (VTU)-RC at Cambridge Institute of Technology, Bengaluru, Karnataka, India. He served for seven years as an Assistant Professor at DVR & Dr HS MIC College of Technology, Kanchikacherla, Krishna Dist., form 2011 to 2018. Now he is currently working as an Assistant Professor in Vikas College of Engineering and Technology, Nunna, Vijayawada, Krishna Dist., AP, India. His areas of Interest are Big Data Analytics, Wireless Networks and Computer Networks.

Dr. K. Satyanarayan Reddy received his M.Sc. & M.Phil. (Mathematics) Degrees from Nagpur University, Maharashtra State, and M. Tech. (CSE with specialization in Computer Applications) from Indian School of Mines [now IIT (ISM)], Dhanbad, Jharkhand in 1987, 1988 and 2000 respectively. He received PhD (Computer Science) degree in the year 2012 from School of Science & Technology, Dept. of Computer Science at Dravidian University, Kuppam, AP, India. He is currently working as Head of the Department of Information Science & Engineering, Cambridge Institute of Technology, Bangalore, Karnataka State, India. His current areas of research are High Speed Networks, Data Communications, Network Security, Wireless Sensor Networks, Big Data, and Artificial Intelligence. Currently he is guiding 7 PhD Scholars under VTU, Belagavi, KN, India. He has 60 publications to his credit in various National & International Journals and International Conferences.

Retrieval Number: D9130118419/2020©BEIESP
DOI:10.35940/ijrte.D9130.018520
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication