A Survey on Big Data Challenges in Fuzzy Algorithms

Kanika Maheshwari
Samrat Ashok Technological Institute
Vidisha, India

Vivek Sharma
Samrat Ashok Technological Institute
Vidisha, India

ABSTRACT
In this paper the survey is done on various challenges that faced during clustering of very large data or fuzzy clustering algorithms that applied over big data in various substantive areas. Big data is a term which is used to define large volume of data. Due to its large size this takes huge volumes to store it thus it is simply inappropriate to use such algorithms that require full data set to analyze data efficiently as data is rapidly increasing and it will require a hard core system to process such larger needs algorithm. In data analysis, clustering plays an important role to find the underlying pattern structure as big data contains so much uncertainty in it, so fuzzy clustering is one of the best methods to capture the uncertainty. In that survey paper we are focusing on the methods and fuzzy algorithms that works well to address fuzzy clustering related problems or challenges**.**

General terms
Clustering, fuzzy set, problems addressing, analysis.

Keywords
Big data, Fuzzy clustering algorithm, Fuzzy challenges**.**

1. INTRODUCTION
Big data is a term which is used to define large volume of data. In today's world use of internet increasing data by day and every single object takes data to process whether it is log data and vice versa, it is estimated that up to 2020 the limit will reach from peta bytes to Exabyte which will take large volumes to store it. As data increasing day by day it become a trouble to store, manage, analyze and process it. As an important technique of data analysis, clustering plays an important role in finding the underlying pattern structure[4].Clustering is a method of finding similar objects that together make a group that is the members of a particular group has most similar properties to each other. Two types of clustering can be performed in order to categorize data for its analysis or insights that is soft or fuzzy clustering and hard clustering. Fuzzy clustering is found to be best method to capture the uncertainty of real data [1]. Many algorithms proposed to perform clustering over big data but very few of them are introduced to address the problem related to fuzzy clustering [2].Conventional clustering means to cluster data into exclusive sets while you know boundaries are imposed to the particular cluster or you can say clearly that the object belongs to the cluster or not [1].In many real world situations such small partition is so insufficient like the condition of getting a customer feedback where various customers have variations in their views like if you ask for a product quality so there would be no straight answer like worst or excellent it would be excellent , very good ,good, bad ,very bad ,worst and vice versa thus fuzzy clustering is use to make cluster for the object having no definite boundaries and it provides degree of membership for objects belongs to the particular cluster. When the degree of membership would be identified, overlapping clusters would be found, but as the technique develop and provide leniency it also become difficult to implement it thus the same with fuzzy clustering, there are very few algorithms present to address the problems related to fuzzy clustering. Fuzzy based clustering often used in Big Data and One of the main problem with handling big data is objects cluster set uncertainty and their overlapping, in condition of cluster overlapping it becomes difficult to store it and analyze it that enhance the complexity of big data handling fuzzy based clustering help here and few methods that work well to produce fuzzy partition include fast FCM (FFCM), multistage random FCM which combines FFCM [2]. while an effective algorithm that address space related problem is kernel fuzzy c-means algorithm (KFCM)[4].

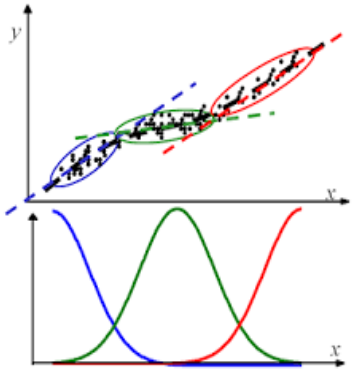

Fig.1: Showing Problems with Fuzzy Based Clustering

2. LITERATURE SURVEY
Dzung Dinh Nguyen, Long T hanh Ngo and Long The Pham [3] proposed a method to handle clustering related problems, thus they introduced Intuitionistic fuzzy sets (IFS) and Intuitionistic type-2 fuzzy sets (InIT2FS) with the motive of handling data uncertainty, by combining the advantages of these fuzzy sets with fuzzy clustering algorithms one can overcome the fuzzy clustering related problem of conventional FCM algorithm.In handling big data as we all know uncertainty of data is the bigger cause which in turn causes its analysis too. In it the uncertainty handling is based on identification of membership functions and identification of non-membership function that is based on an hesitance assessment function [3]The uncertainty and hesitance are two terms different many people mix both as same while there is little difference lies in both the terms. In [3] they introduced Intuitionistic type-2 fuzzy sets (InT2FS) which work on the extension of fuzzy set intuition so it is able to handle uncertainty and hesitance both in the data.While apply that fuzzy sets in any fuzzy clustering algorithm they provide far better results than any traditional algorithm .Thus simply applying Intuitionistic sets rather then simple sets provide better results in clustering uncertain data. Other most popular fuzzy clustering algorithms and applications were introduced in [5], [6], [7].To address problems related to big data like analytics, storage, clustering (also known as unsupervised learning[4].performed literal algorithms that need access to

full data set are inefficient due to memory requirement problems. Thus they introduced a clustering (kernelized fuzzy c-means (KFCM)) technique to address one of the clustering related problems. The literal KFCM algorithm has a memory requirement of O (n2) as per kernel based clustering, where n is the number of objects in the data set. Thus, even data sets that have nearly 1,000,000 objects require terabytes of working [4].So the literal algorithm is infeasible for most of the systems, one way to solve this problem is using incremental algorithm/incremental method in which you perform iteration on small data sets and at last after processing small chunks of data sets results would be combined. Timothy C. Havens, James C. Bezdek, and Marimuthu Palaniswami proposing three kinds of incremental algorithms RSE-KFCM, SP-KFCM, and O-KFCM.They estimate performance of these algorithms by simply comparing their clustering results to the literal KFCM algorithm. They also showed that the following algorithms are producing reasonable partition sets of very large data, by comparing results of all the three algorithms RSE-KFCM found to be most feasible among all as it significantly speed up at low sampling rates too while O-KFCM producing most accurate approximation but have low efficiency [4].Thus to use RSE- KFCM at the highest sample rate allowable for your computational and problem needs [4].

Huber's Description of Data Set Sizes [24, 17]

Bytes 10_2 10_4 10_6 10_8 10_{10} 10_{12} 10>12 ∞
"size" tiny small medium large huge monster **VL** ∞

With every changing day challenges related to big data also changes some of them are new while others are old some more challenges related to big data include its searching, its analysis and its categorization. In [2] Neha Bharill and Aruna Tiwari proposed a fuzzy based supervised classifier for handling searching, storage and categorization related problems of big data. They proposed a Random Sampling Iterative Optimization Fuzzy c-Means (RSIO-FCM) algorithm which effectively handles all that problems by adequately cover all the instances of big data. It has been observed that the classification accuracy of the algorithm is higher than the literal RSE-FCM algorithm. The performance of RSE- FCM and RSIO- FCM that taken over following parameters.

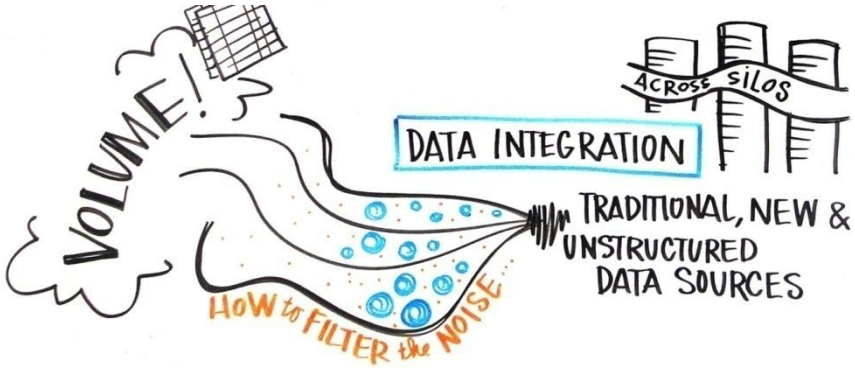

Fig.2: Showing Big Data Nature & Sources

a) The fuzzification parameter (m) that thoroughly affects formation of cluster.

b) The objective of both of the algorithms is the objective function minimization iteratively.

c) The time (t) in seconds is the time taken to group large data and to compute the cluster centers.

d) The accuracy of classification which determines unknown instance searching from very large dataset.

	RSE-FCM		
M	**objective function**	**Total time(s)**	**Accuracy**
1.2	387910619.6	1.54	79.02%
2.3	181503981.8	1.44	72.98%
3.5	79004007.83	1.43	71.09%
4.5	39501984.9	1.36	63.90%

	RSIO-FCM		
M	**objective function**	**Total time(s)**	**Accuracy**
1.2	58137603.98	0.560	87.96%
2.3	12096408.21 0	0.495	91.77%
3.5	772374.015	0.484	92.34 %
4.5	0788.88	0.480	93.22%

Results: N. Bharill and A. Tiwari, Handling Big Data with Fuzzy Based Classification Approach [2].

Table 1&2: Showing performance of both the algorithms

Lun Hu, and Keith C. C. Chan in his paper introduced Fuzzy Clustering Algorithm for Complex Networks(FCAN) algorithm to solve problem appeared in complex networks. As to develop cluster in complex network is a major challenge and most of the real world problems now a day's create complex network where nodes represent object and link between them represent relationship following algorithm helps to solve the problem by focusing on both content and link information[4].

While in [8] Reba Ayeldeen, Aboul Ella Rassanien and Aly Fahm says that the fuzzy logic algorithm poses problems in dealing with large amount of data. In general the main aim of cluster analysis is to collect similar individuals and to classify group on the basis of their feature which is difficult in very large data with simple fuzzy logic algorithm. They proposed Fuzzy Euclidean distance clustering algorithm that helps in partitioning and categorizing very large documents or text into more meaningful categories.

In [9] Yangtao Wang, Lihui Chen, Senior Member, IEEE, and Jian-Ping Mei, Member, IEEE says Clustering is a promising data analysis tool for finding the pattern structure and information underlining unlabeled data.Clustering algorithm that required all the data to store in the memory for analysis purpose become so infeasible when it comes to very large data sets. In [9] they proposed an incremental fuzzy clustering algorithm called incremental multiple medoids-based fuzzy clustering (IMMFC) to manage patterns that are complex and not compact and well separated from each other.

3. CONCLUSION

Handling big data and addressing its challenges is a need now a days, various algorithms are introduced to handle big data by challenges but very few algorithms addressed fuzzy related problems that mostly encountered in very large and unorganized data that is not static. In that paper we introduced some of those algorithms that provide contribution in that way. To handle problems related to big data, fuzzy data techniques and its algorithms should need a boost which will be a great contribution in the IT world ever.

4. REFERENCES

[1] Lun Hu, and Keith C. C. Chan Fuzzy Clustering in A Complex Network Based on Content Relevance and Link Structures, DOI 10.1109/TFUZZ.2015.2460732, IEEE Transactions on Fuzzy Systems.

[2] Neha Bharill and Aruna Tiwari Handling Big Data with Fuzzy Based Classification Approach, DOI: 10.1007/978-3-319-03674-8_21, Springer International Publishing Switzerland 2014

[3] Dzung Dinh Nguyen, Long T hanh Ngo and Long The Pham Interval Type-2 Fuzzy C-means Clustering using Intuitionistic Fuzzy Sets, Third World Congress on Information and Communication Technologies (WICT) 2013.

[4] Timothy C. Havens, James C. Bezdek, and Marimuthu Palaniswami Incremental Kernel Fuzzy c-Means Computational Intelligence, SCI 399 Springer-Verlag Berlin Heidelberg 2012.

[5] H. Frigui and C. Hwang, "Fuzzy clustering and aggregation of relational data with instance-level constraints," IEEE Trans. Fuzzy Syst., vol. 16,no. 6, pp. 15651581, Dec. 2008.

[6] Y. A. Tolias., and S. M. Panas., "Image Segmentation by a Fuzzy Clustering Algorithm using Adaptive Spatially Constrained Membership Functions," IEEE Trans. Syst., Man, Cybern. A,Syst., Humans, vol. 28, no. 3, pp. 359369, May 1998. 263

[7] W. L. Cai., S. C. Chen., and D. Q. Zhang., "Fast and Robust Fuzzy CMeans Clustering Algorithms Incorporating Local Information for image Segmentation," Pattern Recognition, vol. 40, no. 3, pp. 825838, Mar. 2007.

[8] Reba Ayeldeen, Aboul Ella Rassanien and Aly Aly Fahm," Fuzzy clustering and categorization of text documents", 978-1-4799-2439-4113/$3l.00 ©2013 IEEE.

[9] Yangtao Wang, Lihui Chen, Senior Member, IEEE, and Jian-Ping Mei, Member, IEEE, Incremental Fuzzy Clustering With Multiple Medoids for Large Data IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 22, NO. 6, DECEMBER 2014.