REDUCTION OF BIG DATA SETS USING FUZZY CLUSTERING

S.Prabha[1] PG Student, P.Kola Sujatha[2] Assistant Professor
Department of Information Technology, MIT, Anna University

Abstract – **Big Data comprises of large volume, growing data sets from multiple sources. The fundamental requirement is to extract useful information by exploring large volume of data. A preprocessing step of clustering is used to divide data into manageable parts. Fuzzy Clustering adds flexibility for clustering very large datasets in which each object can have membership in more than one cluster. The Incremental Weighted Fuzzy C-Means(IWFCM) introduce weight that describes the importance of each object in the clusters .IWFCM produces cluster with minimum run time and with high quality. The e-book dataset is executed over the Hadoop environment which executes over map reduce framework and data is reduced using IWFCM.**

Keywords – Big Data, IWFCM, Hadoop, Map Reduce

I INTRODUCTION

Big data refers to datasets that cannot be managed with current technologies or data mining software tools due to their large size and complexity. Big data mining is the capability of extracting useful information from these large datasets or streams of data [27].

One reason of Big Data is the growth of the internet where every queries and clicks in google are recorded. On Amazon or eBay, every purchase and every click is captured. When reading a newspaper online, watching videos, or tracking the person finances, our behavior is being recorded.

Big Data processing framework consists of three tiers. Tier I comprises of data accessing and computing, Tier II comprises of data privacy and domain knowledge, Tier III comprises Big Data mining algorithms. We deals with Tier III, where the data mining challenges focusses on algorithm designs which tackles the difficulties raised by the Big Data volumes.

Clustering is a form of data analysis in which data are divided into groups or subsets such that the objects in each group share some similarity. Clustering is used as a preprocessing step to divide data into manageable parts as a knowledge discovery tool, for indexing and compression. The most popular functionality of clustering is to assign labels to unlabeled data (i.e.) data for which no preexisting grouping is known.

Hard clustering refers to the separation of data into distinct clusters, where each and every data element belongs to exactly one cluster. In fuzzy clustering or soft clustering, data elements can participate in more than one cluster, and associated with each element is a set of membership values. These indicate the strength of the connection between that data element and a particular cluster. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) algorithm. The FCM algorithm tries to partition a finite collection of n elements into a collection of c fuzzy clusters with respect to some given criterion.

K-means is one of the unsupervised learning algorithms that solve the well known clustering problem. K-means classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

In the FCM approach, instead, the same given data does not belong exclusively to a well defined cluster, but the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.

A matrix U is formed whose factors are the ones taken from the membership functions .The number of rows and columns represents how many data and clusters are considered. Two ways to cluster the large data sets are the distributed clustering which is based on various incremental styles, and the clustering a sample found by either progressive or random sampling.

Weighted Fuzzy C-Means(WFCM) introduces weights that define the relative importance of each object in the clustering algorithm.

II RELATED WORK

The Big Data Analytics has been divided into three tiers [28]. The first tier deals with the data accessing and computing second tier deals with the privacy considerations and the third tier deals with the data mining algorithms. The main problem in data mining is to generate global models by combining locally discovered patterns to form a unifying view.

The Big Data is growing extremely, for high dimensional data the data reduction is important. The medical datasets are of high dimensionality in each field. The data reduction is easier for non-densed data rather than dense datasets [29].

The large datasets can be done by combining clustering and classification [2]. To obtain robust and stable clustering, consensus functions can be applied for clustering ensembles combining a multitude of independent initial clusterings. Direct applications of consensus functions to highly dimensional data sets remain computationally expensive and impracticable. Therefore a multistage scheme including various procedures for dimensionality reduction, consensus clustering of randomized samples, followed by the use of a fast supervised classification algorithm is needed.

The ant colony optimization technique has emerged as a novel meta-heuristic belongs to the class of problem-solving strategies derived from natural (other categories include neural networks, simulated annealing, and evolutionary algorithms) [8]. The ant system optimization algorithms is basically a multi-agent system where low level interactions between single agents (i.e., artificial ants) result in a complex behavior of the whole ant colony..

A hybrid form of clustering which combines one or more clustering techniques can be used to cluster very large medical datasets [30].The four kinds of hybrid fuzzy cluster ensemble frameworks are used to cluster bio-molecular datasets.

For preserving privacy in data-intensive applications, Twice- privacy algorithm based on utility matrix and multiattribute clustering had been used [23]. Twice –privacy conducts a clustering of sensitive values to protect similarity, sets different weight to retain quasi-identifier attribute to query service. In cloud environments distributed clustering which includes a novel distributed high dimensional data clustering algorithm based on Map-Reduce framework to distinguish the different communities from the entire social network had been suggested [16].

K-Means algorithm had been proved to be better than FCM algorithm [25]. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. In fact, FCM clustering which constitute the oldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster.

Single Pass (through the data) Fuzzy C-Means algorithm which is based on Weighted Fuzzy C-Means neither uses any complicated data structure nor any complicated data compression techniques, yet produces data partitions comparable to Fuzzy C-Means[14]. Simple Single Pass Fuzzy C- Means clustering algorithm when compared to Fuzzy C-Means produces excellent speed-ups in clustering and thus can be used even if the data can be fully loaded in memory.

Clustering technique on uncertain data (ie) clustering uncertain objects with the uncertainty regions defined by pdfs was difficult [26]. For an accurate representation, at least thousands of sample points should be used to approximate an object's pdf. When applying the UK-means algorithm to cluster uncertain objects, a large number of expected distances have to be calculated. The basic min-max-dist pruning method is fairly effective in pruning expected distance computations.

For finding the number of fuzzy clusters a new cluster validity index fwith crisp and fuzzy data had been suggested [9]. The new index, called the ECAS-index, contains exponential compactness and separation measures. These measures indicate homogeneity within clusters and heterogeneity between clusters, respectively. Moreover, a fuzzy c-mean algorithm is used for fuzzy clustering with crisp data, and a fuzzy k-numbers clustering is used for clustering with fuzzy data.

Both weight based Fuzzy C-Means algorithm like Single Pass and Online Fuzzy C-Means can be converged in clustering the image datasets[14],[15].Both algorithms weight ex-amples and cluster subsets of weighted examples.

Bit reduced FCM had been used to reduce the number of distinct patterns which must be clustered without adversely affecting partition quality [24]. The reduction was done by aggregating similar examples and then using a weighted exemplar in the clustering process. The reduction in the amount of clustering data allows a partition of the data to be produced faster.

III PROPOSED SYSTEM DESIGN

The fuzzy clustering proves to perform better for noisy data. Since the sources of Big Data are majority the social networks fuzzy clustering is well for data reduction .The data reduction helps in easy storage and retrieval of large data and for extracting useful information. Incremental fuzzy clustering is used to improve performance in very large data sets. Reduction of is highly concentrated. High quality clustering is produced and the storage and retrieval of data is made easier.

Fig.1 shows the architecture which describes the working of IWFM.

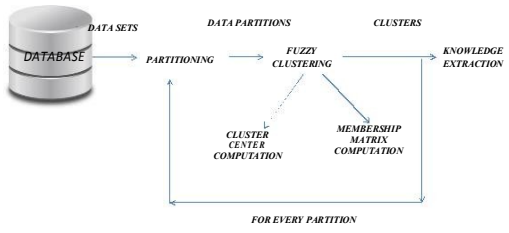

Fig.1 IWFCM System Architecture

The datasets are being partitioned and each partitioned undergoes cluster center calculation and membership values. Useful knowledge can be extracted from the result.

IV IMPLEMENTATION

The following algorithm is implemented over the hadoop environment which follows map reduce framework for large datasets. The datasets are map reduced and the output is further implemented by IWFCM over the environment.

Algorithm :IWFCM

Input - X,c,m,n_s
Output - V
1: Set the weight vector (w) for the first subset as 1
2:Calculate partition matrix (U) and cluster centroid(V) for the first subset using

$$U_{ij}=[\sum(\|x_j'-v_i\|/\|x_j - v_i\|)^{2/m-1}]^{-1 \text{ for all } i,j}$$

$$V_i=(\sum w_j (u_{ij})^m x_j)/(\sum w_j(u_{ij})^m) \text{ for all i For subset 2 to s}$$

3:Calculate weight $w_i = \sum(u_{ij})w_j$, i=1….c, j=1…n_s 4:Set w=w'
5: Calculate wFCM({VUX_l},c,m,w,V)

X - Input Dataset
c - No. of Clusters
m - Fuzzification Constant
n_s - No. of Subsets

Fig.2 Algorithm of IWFCM

Fig.2 gives the algorithm to produce clusters using IWFCM.

V RESULTS

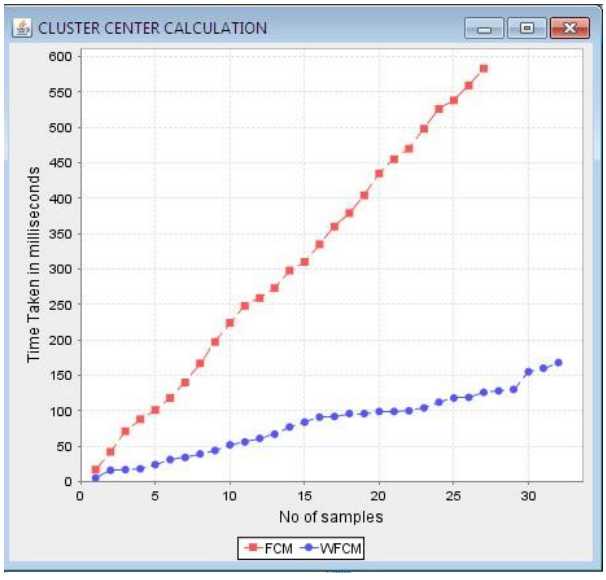

Fig.3 Comparison Graph

In Fig.3,the time taken by FCM and IWFCM for each iterations have been compared . IWFCM proves to be good since weight is being added to each dataset.

VI CONCLUSION AND FUTURE WORK

An effective way to cluster the large volume of data is IWFCM by extending the clustering of sampled data to loadable and unloadable datasets. The Distributed Environment has been set up where the very large datasets need to be reduced. The incremental fuzzy clustering can be enhanced with adding indexing like bitmap indexing and aggressive indexing applied to the distributed environment like hadoop for easy retrieval of information. The run time can be reduced due to indexing.

REFERENCES

[1] Anderson.D, J. Bezdek, M. Popescu, and J. Keller, "Comparing fuzzy,probabilistic, and possibilistic partitions," IEEE Transaction on Fuzzy Systems, vol. 18.Issue 5,October 2010.

[2] Andrei Kelarev, Andrew Stranieri, John Yearwood, Herbert Jelinek "Empirical Investigation of Consensus Clustering for Large ECG Data Sets", International Conference on Pattern Recognition, Informatics and Medical Engineering 2012.

[3] Belabbas. M and P. Wolfe, "Spectral methods in machine learning and new strategies for very large datasets," Proceedings National Academic Science U.S.A., vol. 106,no. 2, pp. 369–374, 2009.

[4] Bezdek.J and R. Hathaway, "Convergence of alternating optmization," Neural, Parallel, Science Computing, vol. 11, no. 4, pp. 351–368, Dec. 2003.

[5] Bo.W and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in Proceedings International Conference on Computer Vision, pp. 1–8,Oct.2007.

[6] Boyd, danah and Kate Crawford. (2012). "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon."Information, Communication, & Society 15:5, p. 662-679.

[7] Callaghan.L, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality clustering," inProc. IEEE International Conference on Data Engineering, Mar. 2002, pp. 685–694.

[8] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai "A New Data Clustering Approach for Data Mining in Large Databases" International conference on Parallel Architectures 2002.

[9] Fazel Zarand, M.R. Faraji and M. Karbasian," An Exponential Cluster Validity Index for Fuzzy Clustering with Crisp and Fuzzy Data", Transaction E: Industrial Engineering Vol. 17, No. 2, pp. 95 Sharif University of Technology, December 2010.

[10] Frigui.H,"Simultaneous clustering and feature discrimination with ap-plications,"in Advances in Fuzzy Clustering and Feature Discrimination With Applications. New York: Wiley, 2007, pp. 285–312.

[11] Guha.S, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan,"Clustering data streams: Theory and practice," IEEE Transaction on Knowledge and Data Engineering, vol. 15, no. 3, pp. 515–528, May/Jun. 2003.

[12] Gunnemann.S, H. Kremer, D. Lenhard, and T. Seidl, "Subspace clustering for indexing high dimensional data: A main memory index based on local reductions and individual multi-representations," in Proceedings International Conference, Uppsala, Sweden, 2011, pp. 237–248.

[13] Hathaway.R and J. Bezdek, "Extending fuzzy and probabilistic clustering to very large data sets," Computational Statistics and Data Analysis, vol. 51,Issue 1, November 2006,pages 215-234.

[14] Hore.P, L. Hall, and D. Goldgof, "Single pass fuzzy c means," in Proc.IEEE International Conference on Fuzzy Systems, London, England, 2007.

[15] Hore.p, L. O. Hall, and D. B. Goldgof, "A fuzzy c means variant for clustering evolving data streams," in Proceedings IEEE International Conference System, Man, Cybern., Oct. 2007, pp. 360–365.

[16] Hui Liu, Wu Qu, Jin Yi, Junhe Wang ,Chenghao Sun," A distributed clustering method to segment micro-blog users on cloud environments" , 2013 IEEE on Fuzzy Systems, London, England, 2008.

[17] Jiawei Han,Micheline Kamber,Jian Pei "Data mining Conepts and Techniques" Third Edition book,2009.

[18] Keerthika Janani.M ,Sudhakar.G," HAVS: Hadoop Based Adaptive Video Streaming by the Integration of Cloudlets and Stratus", Volume 2, Issue 3, February 2014 ,Page No 408-413, IJARCET

[19] Lawrence O. Hall, and Dmitry B. Goldgof, "Convergence of the Single-Pass and Online Fuzzy C-Means Algorithms" IEEE Transactions on fuzzy systems, vol. 19, no. 4, august 2011.

[20] Liran Einav and Jonathan Levin, "The data revolution and economic analysis" NBER Innovation Policy, Economy Conference, April 2013.

[21] Mathew Smith,Christian Szongott, Benjamin Heme, Gabriele von voigt ," Big Data Privacy Issues in Public Social Media" IEEE International Conference on Data Engineering, Mar 2013.

[22] Obu-Cann.K, K. Iwamoto, H. Tokutaka and K. Fujimura, "Clustering by SOM (self-organising maps), MST (minimal spanning tree) and MCP (modified counter-propagation)," 6th IEEE International Conference on Neural Information Processing, pp. 986-991, vol. 3, 199.

[23] Qing Zhu and Ning Li ,"Privacy Protection by Multiattribute Clustering in Data- intensive Service" IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications.

[24] Ramesh.N, Anil.D, Kiran.M." Securing Distributed Accountability for Data Sharing in Cloud Computing", Volume 2, Issue 8,August 2013, Page No.2514-2519, IJARCET

[25] Soumi ghosh and sanjay kumar dubey," comparative analysis of k-means and fuzzy c-means algorithms" , (ijacsa) International Journal of Advanced Computer Science and Applications ,vol. 4, no.4, 2013.

[26] Wang Kay Ngai, Ben Kao, Chun Kit Chui," Efficient Clustering of Uncertain Data", Proceedings of the Sixth International Conference on Data Mining 2006.

[27] Wei Fan and Albert Bifet, "Mining Big Data:Current Status, and forecast to the future" http:// big-data-mining.org/.

[28] Xindong Wu, Xingquan Zhu ,Gong-Qing Wu "Data Mining with Big Data" IEEE Transactions on knowledge and Data Engineering 2013.

[29] Xue-min.A MAO, B. Chuan-xi CAI, andC. Bing-yu SUN "Comparative Research on Methods of Dimensionality Reduction in High-dimension Medical Data" Fourth International Workshop on Advanced Computational Intelligence; 2011.

