International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Overview of Fuzzy Clustering Algorithm on Hadoop Map Reduce Computing

Mrs. C.Sunitha [1], Ms. I.Jeevitha [2]
[1] Professor & Head of the Department, [2] Assistant Professor
Department of BCA & M.Sc.SS
Sri Krishna Arts and Science College

ABSTRACT

Big Data is a one vital component in today's era. It deals with large amount of heterogeneous data in every second. It is very difficult to handle the data using traditional data processing applications. Big Data is all about processing and tuning the complex, invaluable and unstructured data into well structured information. Very huge quantity of data about text, image and video are produced day by day, where it is difficult to process particular data sets from the gathered collection of data. Clustering is an essential technique in big data mining. Clustering is the process of grouping the data based on their similar property. It groups input data sets into similar subsets, is called as clusters. A software framework called Hadoop helps distributed processing of huge data across distributed computers and clusters with the help of Map Reduce Programming. Map reduce Programming solve the problems that faces during the computing the clusters. In this Paper we present a review of fuzzy K-mean Clustering algorithm to process data which help researcher to decide the efficient algorithm for data processing.

Keywords: Big Data, Map Reduce clustering, fuzzy C Means Clustering, parallel computing.

I.INTRODUCTION

In recent era the usage of internet plays a vital role, every day the internet user's access data from various sources as text file, image file, audio file and video files. The extraction of unique data is not providing limited boundary where it tend to different collection of data sets. It is great challenge to the user accessing, searching, and storing the heterogeneous type of data[9]. There is need of new application which minimizes time to access data by avoiding the loss of data too. These large unstructured data cannot simply used for further processing. The complex data also known as multi structure or multi source data. To cluster this heterogeneous, diverse and autonomous data source, system uses map reduce technique. Clustering is process of analyzing the data and process useful information. Where data similar are grouped in a single cluster, data objects in same cluster have same properties, data object from different cluster posses different characteristic. We can easily find the dense and sparse regions and easily search corresponding pattern and data attributes. Clustering need to process several tasks, list out steps.

(i) Pattern Representation
(ii) Measurement appropriate data domain
(iii) Clustering or grouping
(iv) Data abstraction
(v) Assessment of output

First the patterns are identified based on the selection process in which it identify the most similarities in the existing pattern and derive a new prominent set [5]. Second step is to cluster the data with appropriate metrics. The next step the clustering algorithm will be applied to the prominent selected sets. The fourth steps define the prototype for the clustering extracting prominent compact representation of data sets. The Last step produces the cluster results any cluster in nature. The evaluation of cluster involves several aspects: cluster tendency, cluster validity. Three types of validation process exist they are internal, external and relative. Cluster validity is process of evaluating clusters , internal validity check the structure is proper for data , external validity compare the recovered structure to prior one and the relative validity is to compare two structures

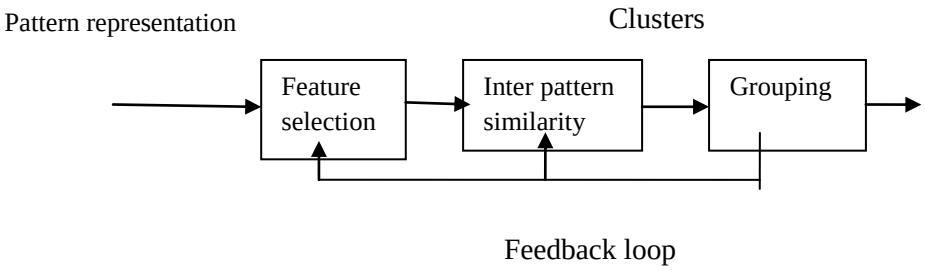

Figure 1: Clustering Procedure

In figure it represent the process of the clustering first perform the feature selection and extraction of the data. Second step it perform the selection of the appropriate clustering algorithm and the last step perform the evaluation of the cluster output. The K-means algorithm have problem of producing hard partitioning of the data set it means only one cluster is depicted at a time. The data points on the adjacent cluster or edge of the cluster may not be in the cluster set. The fuzzy K-means algorithm also called as Fuzzy C-Means algorithm introduced by Bezdek , where at each point has the probability of dependency in each cluster [6] . The coefficient value is associated with every point gives the degree of the k_{th} cluster and sum of coefficient values should be one. The increasing capacity of dataset, the main memory is not providing a sufficient environment for processing the data. The solution to the problem is Apache Hadoop is able to process very huge data sets. With the help of MapReduce , Hadoop extract the data from large datasets by processing the query. The machine learning library in Hadoop called Mahout where it produces scalable performance for Hadoop [9].

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

II. METHODOLOGY

Clustering is an unsupervised learning concept and aggregates the similar data sets in different classes. The classes are referred as clusters. Samples inside the class of high similarity compared to the other classes of different samples. This technique used in many areas like data mining, statics and machine learning. Fuzzy K-mean clustering technique is based on centroid clustering technique [6].

A. Hadoop Platform

Hadoop is an open source framework; it includes two components namely Map Reduce and Hadoop Distributed File System (HDFS). The implementation of the map reduce need these two components. The large heterogeneous data sets are stored in HDFS, using MapReduce we need to process the stored data in HDFS. The retrieved files are split into contiguous chucks of each 64MB by default. Each of these is simulated in different racks. HDFS has two nodes; First name node stores the metadata and the data nodes stores the blocks from file. The name node and data node in similar refer as jobtracker and Tasktracker. The jobtracker assign the jobs to the tasktracker. Which then process the job using MapReduce model.

B. MapReduce Technique

There are two main programs associated with MapReduce first is Map and Reduce. According to the Hadoop block size dataset will be split for processing. Map() function generates <key, value> intermediate set by associating with each block. The function Reduce() aggregates the intermediate result and produce the final output. Like MapReduce in HDFS, Hadoop also have master/ slave architecture.

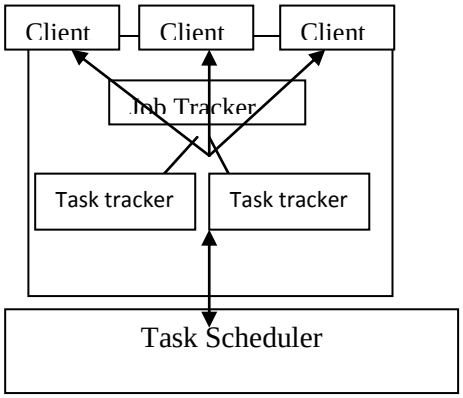

Figure 2: Architecture of Map Reduce of Hadoop

MapReduce linearly order the distributed operation on data sets based on key/value pairs. MapReduce works composed by Tasktracker and Jobtracker. MapReduce works in two phases first a map phase and second reduce phase. In Map phase the input data split into a several number of fragments and each fragment is allocated to map task. In Map phase, it splits the input data set into a larger number of fragments and allocates it to map task. The framework then distributes the map tasks across many clusters of nodes [7]. Each map task process the key/value pairs and produces intermediate key/value pairs. The input key / value pair (k ,v) transmit into input into diverse pair

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

(k^1,v^1) by user defined function invokes by map task. After processing the intermediate set it produce the set of tuples (k^1,v^1) , so that specific key appear together. It also fragment set of tuples into number of reduce tasks. For each tuples it calls the user defined function and transmits the tuples into output key. After generating the key values in reduce, the reduce task distributed across the cluster of nodes. If any node fails in the middle of computation the task redistributed among the outstanding nodes. Performing many map and reduce tasks increase good load balancing and fault tolerance. Figure 3shows the MapReduce Processing Model [8].

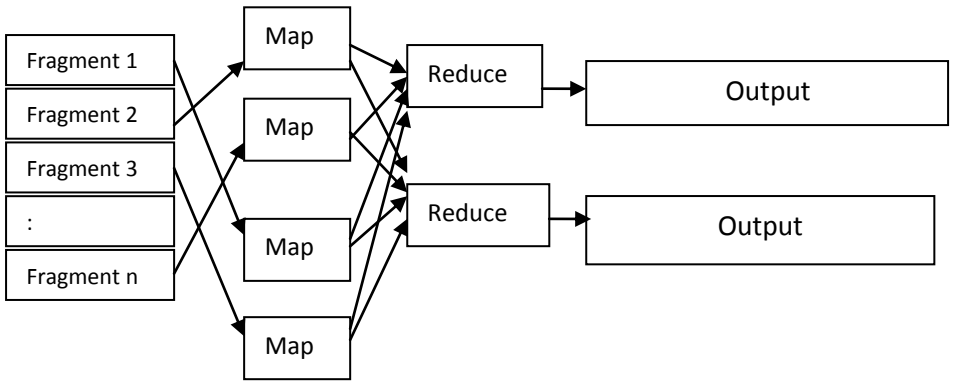

Figure 3: Map Reduce Computing Model

III. FUZZY K-MEAN CLUSTERING

Fuzzy k-mean clustering also called as soft clustering technique it is an expansion of K-mean clustering. It reduce the intra cluster variation series. Bezdek introduce the fuzziness parameter (M) in fuzzy k-mean which determine the degree of fuzziness in clusters [2]. The Fuzzy k-mean algorithm is as follows:

1. select number of clusters
2. Computes distance matrix from xj point , where each cluster centers consider as Euclidean distance between the point , cluster using [3].

$$dij = \sqrt{\sum(\mathrm{xj} - \mathrm{ci})2}$$

Where Euclidean Distance between data point from j_{th} and i_{th} cluster is equal to d_{ij}.

3. The membership matrix is created with $\mu(x_j)$
4. Then new centroid for every cluster is created based on the previous $C_{i.}$

Stopping criteria: The algorithm will continue until any cluster center point not changed beyond the convergence thresholds and neither the points changes in cluster. In large dataset it is difficult to compute, if overlapping cluster is large i.e. Number of iteration increase tends to increase the execution time. To solve this problem, MapReduce approach is used [1].

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

IV. MAPREDUCE APPROACH

This approach partitions the dataset and then computes on the partitioned dataset also known as job computation. In parallel where individual data set are processed by map and sorted output from the maps.

Inputs : Number of cluster , randomly selected centroid access point , Data point

Output : processed Final centroid and clustered points

- ***Map Algorithm :***

1. Centroid points which randomly selected consider as key and vector .
2. Compute Euclidean distance between vector point and centroid point
3. Calculate the association value of each vector point, create member ship matrix using step 2.
4. Using nearest centroid clusters are generated and assign the data points to the cluster.
5. Maintain the details of vector point and cluster holding it.

- ***Reduce Algorithm:***

1. Each cluster centroid will be recalculated.
   The calculated centroid would go linearly to map and then it iterating . It computes total number of reducers are less than total number of mappers (M<N)

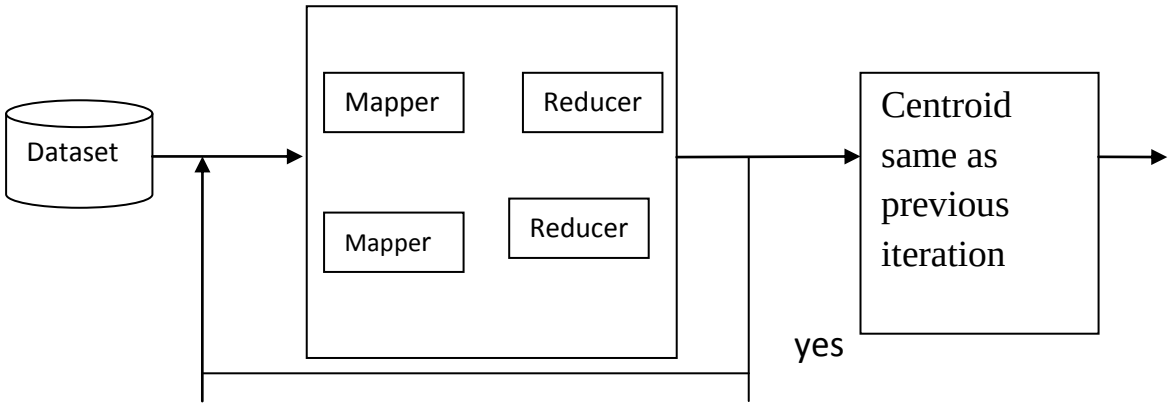

Figure 4: Fuzzy k-mean clustering Algorithm in MapReduce

The figure 4 depicts the function of Fuzzy k-mean clustering computation with the clustered data sets. The Map and Reduce Perform the calculation until the centroid iteration reduce to the minimum values [4].

V.CONCLUSION

In this paper we discussed about the parallel fuzzy algorithm on MapReduce computing environment of Hadoop. The main issues and process in parallel fuzzy k-means algorithm has discussed. The rise of the cloud computing, the research work of clustering and data mining gradually becomes a hot concept of scholars. Study of parallel clustering algorithm tend to the improve the clustering performance, various clustering technique performance are lead to the development of good computation, which process data set quickly. The security and privacy issues on cloud computing are to be concern. Solving the problem will play a key role in cloud computing application in current business areas.

REFERENCES

[1] . Jerril Mathson Mathew, Lekshmy P Chandran "Parallel mplementation of Fuzzy Clustering Algorithm Based on MapReduce Computing Model of Hadoop –A Detailed Survey" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4740-4744.

[2] Bezdek, James C, "FCM : THE FUZZY c-MEANS CLUSTERING ALGORITHM", vol. 10, pp191-203, 1984.

[3] . J. P. Mei and L. H. Chen, "A fuzzy approach for multitype relational data clustering," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 2, pp. 358–371,Apr. 2012.

[4] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Multiple kernel fuzzy clustering,"*IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 120–134, Feb. 2012.

[5] . keshav Sanse , Meena Sharma "Clustering Methods for Big Data Analytics ", International journal of advanced Research in computer Engineering & Technology (IJARCET), Vol 4, Issue 3 , March 2015.

[6] .Garima Sehgal, Dr. Kanwal Garg "Comparison of Various Clustering Algorithms" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3074-3076.

[7] Vaishali R. Patel1 and Rupa G. Mehta "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011.

[8] . Madhavi Vaidya "Parallel Processing of Cluster by Map Reduce", International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.1, January 2012.

[9] .C.Sunitha , I.Jeevitha " A Review of Genetic Algorithm Practice in Hadoop Map Reduce " , International Journal of Science Technology & Engineering Vol 2, Issue 5, ISSN: 2349-784X