International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

SBKMMA: Sorting Based K Means and Median Based Clustering Algorithm Using Multi Machine Technique for Big Data

E. Mahima Jane[a*], Dr. E. George Dharma Prakash Raj[b]

[a]*Asst. Prof., Department of Computer Application , Madras Christian College, Tambaram – 600 059*

[b]*Asst. Prof., Department of Computer Science and Engineering, Bharathidasan University, Trichy - 620 023*

[a]*Email: mahima.jane@gmail.com* ,[b]*Email: georgeprakashraj@yahoo.com*

Abstract

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. KMeans is a traditional partition algorithm which is simple and popularly used. This algorithm has disadvantages such as to identify K clusters, initial allocation etc. In this paper we mainly focus on the initial centroids and improving the efficiency by reducing the number of iterations. Sorting based KMeans algorithm and Sorting based KMedian algorithm are enhanced form of KMeans algorithm where the data are sorted and uses KMeans algorithm. The proposed algorithm focuses on the initial centroid selection with the help of sorting. Here the centroids are default assigned to the objects in the beginning after sorting.

Keywords: KMeans;BigData; Clustering; Sorting.

l. Introduction

Big data analytics refers to the strategy of analyzing large volumes of data, or big data. Big data is a term that refers to huge repository of datasets that is hard and complex to process with the conventional data processing systems. Numerous challenges are in place with big data like storage, transition, visualization, searching, analysis, security and privacy violations and sharing.[1] The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it.

---

* Corresponding author.

A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Partitioning method creates k partitions (clusters) of the known dataset, where all partitions represent a cluster. And each cluster can be represented by a centroid or a cluster representative which is some kind of summary explanation of all the objects present in a cluster[2]. MapReduce is a computational approach that involves breaking large volumes of data down into smaller batches, and processing them separately. MapReduce is a programming model, Google has used successfully is processing its big data sets [2]. Traditional KMeans algorithm centroids. SBKMA and SBKMEDA uses sorting and mean , median for initial centroid. This paper proposes an algorithm by eliminating the drawback of initial centroid assignment of the traditional KMeans algorithm, SBKMA algorithm and SBKMEDA algorithm by default centroids the the objects.

This paper is further written as follows. It discusses other related and relevant work in Section 2, Section 3 gives a small introduction about Sorting Based Machine Clustering Technique followed by the proposed work in section 4.The Experimentation and Analysis is explained in Section 5 followed by the Conclusion in Section6.

2. Related Work

Mahima Jane and his colleagues [3] proposes an algorithm called SBKMA for improving the K-Means algorithm for numeric data. In the SBKMA algorithm, the centroids are identified by sorting the objects first and then identifying the mean from the partition done as per the K clusters.

Each K clusters are partioned and mean of each cluster is taken as centroid. Mahima Jane and his colleagues [4] propose an algorithm called SBKMEDA by improving the SBKMA by taking median as centroids. Multi-Machine clustering technique allow to breakdown the huge amount of data into smaller pieces which can be loaded on different machines and then uses processing power of these machines to solve the problem. Hence number of iterations and Execution Time are considerably reduced.

Anand M. Baswade and his colleagues [5] in this paper the drawback of traditional K-Means algorithm of selecting initial centroid is removed. The enhanced method of the k-means algorithm includes the computation of the average of objects to improve the centroids initialization. Aleta and his colleagues [6] shows the comparison between K-Means and the proposed K-Means algorithm, and it proves that the new method of selecting initial centroids is better in terms of mathematical computation and reliability. The proposed k- means algorithm by Kaur and his colleagues [7] solve the problem of dead unit and optimizes the selection of initial centroids of clusters by using most populated area as a centroid of cluster. Rajalalskmi and his colleagues [8] performed analysis on kmeans algorithm for disease prediction.

3. Multi Machine Clustering Technique

Multi Machine Clustering technique is a clustering method where data are analyzed using multiple machines. It enables to reduce the execution time and improves the speed of the clusters formed. Multi machine clustering technique is divided into parallel clustering and map reduced based clustering. This type of techniques allows the data to be divided into multiple servers and processes the request.

4. Sorting based K Means and Median based Clustering Algorithm using Multi Machine Technique for Big Data

Hadoop is one of most popularly used open-source data analysis tool. It is an implementation of MapReduce for the analysis of large datasets. Hadoop uses a distributed user-level file system, to across the cluster. The file system is called HDFS, and is written in Java. It is designed for portability across heterogeneous hardware and software platforms[1].

In the proposed algorithm Hadoop MapReduce is used to sort the data using multiple machine. After sorting the data are d1ded into K clusters. The average of the mean and median of each group are calculated and fixed as the initiail centroid for the objects in the groups formed. Calculate the distance between the centroid and the objects. Calculate the distance between the centroid calculation using mean till there is no change in the formation of the group. Finally when there is no change in the cluster formation the execution is stopped.

This algorithm focuses to remove the drawbacks of the existing SBKMA and SBKMEDA. In SBKMA sorting is performed in the beginning via parallel processing using hadoop mapreduce. The SBKMEDA uses the same concept of SBKMA but median is used instead of mean.The proposed algorithm first sorts the data using hadoop map reduce and then divides the data into k clusters. Mean and Median are taken from each group and the average value is taken as centroids for the objects in the group.

4.1. SBKMMA Algorithm

Step1 : Start the Program

Step2 : Load the dataset

Step3 : Sort the data

Step 4: Input k clusters

Step5: Divide the data into k clusters

Step6: Calculate the average of mean and median of each group.

Step7: Mark them as initial centroids to the objects present in the group

Step8: compare the objects with other centroids

Step9: Group the objects with minimum distance

Step10 : Take mean and continue step 8 and 9till there is no change in the formation

Step 11 : Stop

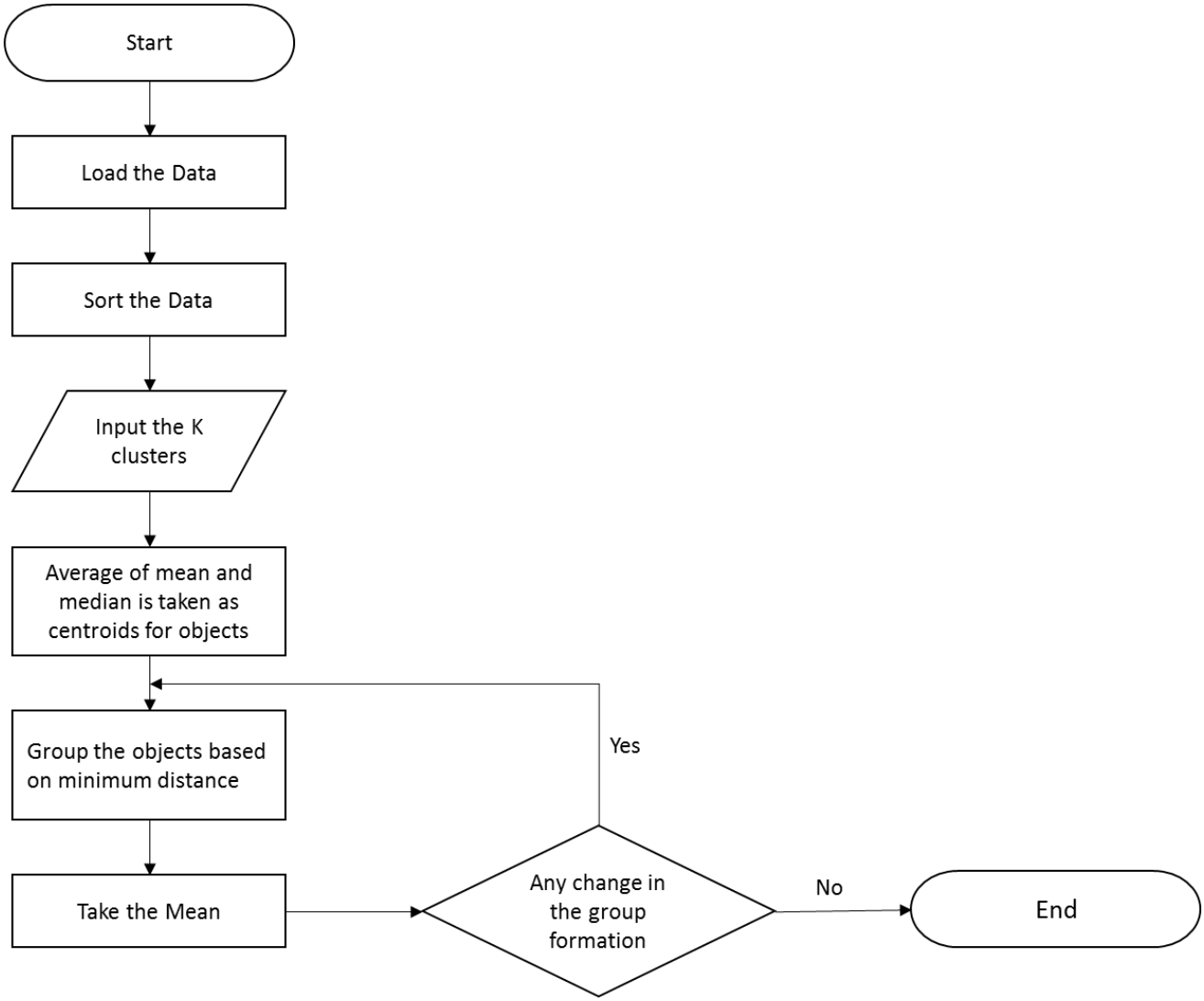

Figure 1: Working of the SBKMMA Algorithm

5. Experimentation and Analysis

SBKMA algorithm, SBKMEDA algorithm and SBKMMA Clustering Algorithms are implemented in HadoopMap Reduce framework. Hadoop Map Reduce is a software framework for easily writing applications which process vast amounts of data in- parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner [1]. Sorting is performed using Hadoop MapReduce.

Map Reduce usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.[3].

The process of Experimentation of this paper using MapReduce is explained next. A randomly generated file containing mobile dataset of 1 Terabyte is generated. This data is stored in the form of text files.

Table 1: Execution Time

No. of Nodes	Clustering Algorithm – SBKMA	Clustering Algorithm - SBKMEDA	Clustering algorithm – SBKMMA
5 – Nodes	20.56	18.5	15
10 –Nodes	16.46	14.3	13.2
15 – Nodes	9.83	8.1	7.7
20 – Nodes	4.56	4	3.8

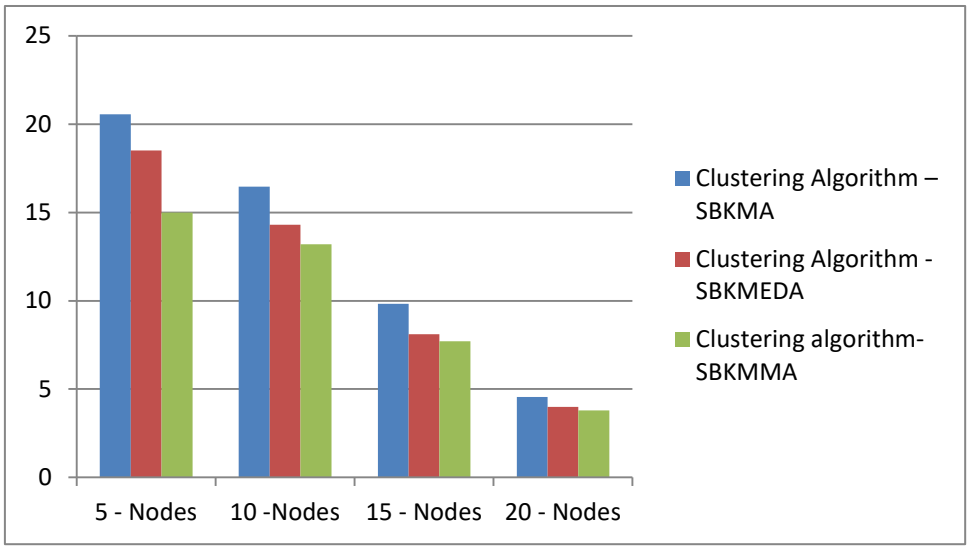

Figure 2: Execution Time

The execution time of SBKMMA is lesser when compared to SBKMA and the SBKMEDA algorithm. When the number of nodes are increased the efficiency increases even more . This can be seen in fig2. Execution time.Execution time is measured in seconds. The number of nodes are different when compared to the previous SBKMA algorithm.

Table 2: No of Iterations

No. of Nodes	Clustering Algorithm - SBKMA	Clustering Algorithm - SBKMEDA	Clustering Algorithm - SBKMMA
5 – Nodes	16	12	6
10 –Nodes	12	9	4
15 – Nodes	8	6	3
20 – Nodes	4	3	2

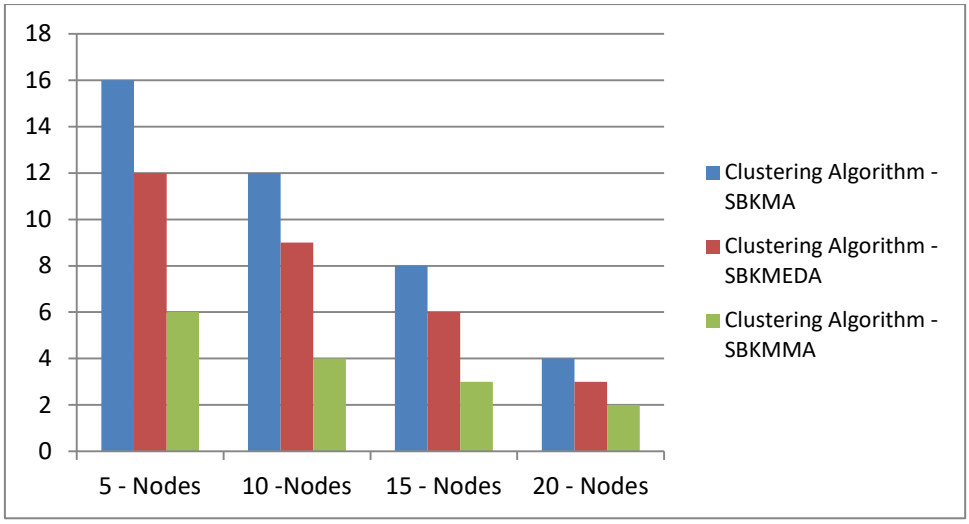

Figure 3: No. of Iterations

Here the number of times the algorithm iterated is less when compared to SBKMA and the SBKMEDA as shown in Fig.3.

6. Conclusion

The proposed Sorting based K Means and Median based Clustering Algorithm for Big Data is examined with the already existing Sorting based KMeans algorithm and Sorting Based KMedian Algorithm. Here multi machine technique with hadoop map reduce is used to sort the data. To improve the efficiency and the average of mean and median are taken as centroids for individual clusters. As the initial centroids are taken based on the k clusters their mean and median reduces the complexity of the iteration. Hence the number of iterations and the time taken reduces more when compared to SBKMA and SBKMEDA. One drawback of the proposed algorithm is that the value of k, the number of clusters, be given by the use regardless of the data. In the future with the help of statistical tool data can be analysed and k can be fixed.

References

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

[2] Nirali Honest and Atul Patel A SURVEY OF BIG DATA ANALYTICS, International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016

[3] Mahima Jane and Dr. E. George Dharma Prakash Raj “SBKMA : Sorting based K- Means Clustering Algorithm using Multi Machine Technique for Big Data “ in the International Journal of Control Theory and Applications Volume 8 2015pp 2105- 2110

[4] Mahima Jane and Dr. E. George Dharma Prakash Raj “SBKMEDA : Sorting based K- Median

Clustering Algorithm using Multi Machine Technique for Big Data " accepted for Advances in Intelligent Systems and Computing.

[5]. Anand M. Baswade, Prakash S. Nalwade2,"Selection of initial centroids for K-Means Algorithm" IJCSMC, Vol. 2, Issue. 7, July 2013, pg.161 –164

[6]. Aleta C. Fabregas, Bobby D. Gerardo, Bartolome T. Tanguilig III,"Enhanced Initial Centroids for K-means Algorithm", International Journal of Information Technology and Computer Science(IJITCS), Vol.9, No.1, pp.26-33, 2017. DOI: 10.5815/ijitcs.2017.01.04

[7] Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013.

[8]K.Rajalakshmi,, Dr.S.S.Dhenakaran,N.Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015