I J C T A, 8(5), 2015, pp. 2105-2110
© International Science Press

SBKMA: Sorting based K-Means Clustering Algorithm using Multi Machine Technique for Big Data

E. Mahima Jane* and E. George Dharma Prakash Raj**

Abstract: Multi Machine Clustering technique combined with K-Means clustering algorithm is one of the very efficient methodused in the Big Data to mine and analyse the data for insights. One of the main disadvantage of K-means clustering algorithms is the deficiency in randomly identifyingthe K number of clusters and centroids. This results in more number of iterations and increased execution times to arrive at the optimal centroid. As a result, the number of iterations and the execution time increases due to multiple iterationin order to arrive at the optimal centroid, and due to a convergence in arriving at the cluster. This paper discusses a modified algorithm to overcome the above said inefficiencies. SBKMA: Sorting based K-Means Clustering Algorithm using Multi Machine Technique introduced in this paper helps reduce the number of iterations and decreases the execution time significantly

Keywords: Big Data, clustering, K-means algorithm, Hadoop MapReduce.

1. INTRODUCTION

The need for Big Data Analysis has tremendously increased in the recent years. Many social networking websites such as Facebook, Twitter havebillions of users who produce gigabytes of contents per minute, similarly many online retail stores conduct business worth millions of dollars. Hence it is necessary to have efficient tools to analyse and group the data and derive meaningful information[1]. Clustering is one such technique which helps derive meaningful insights from Big Data. This technique isthe taskof grouping the data such that the objects in same group are similar to each other than those in the othergroup. In the recent times multiple machine clustering techniques has gained significant attention because theyhave proved to be faster and scalable. K-means is a very powerful partition clustering algorithm because it is a very simple, statistical and quite scalable method. K-means algorithm implemented on Apache Hadoop distributed across multiple machines has become a viable and proven method for clustering Big Data.

In this paper, we have proposed a modified algorithm SBKMA: Sorting based K-Means Clustering Algorithm using Multi Machine Techniqueto the traditional K-Mean algorithmto help overcome inefficiencies in the traditional method.. As a result there is a reduction in thenumber of iterations thereby improving the efficiency to form the clusters in less time.This paper elaborates the proposed SBKMA algorithm further. It discusses other related and relevant work in section II, clustering techniques with Hadoop are mentioned in section III, the proposed SBKMA algorithm is discussed in section IV and results are illustrated in section V.

2. RELATED WORK

Yaminee S. Patil et al., [2] presents the implementation of K-Means Clustering Algorithm over a distributed environment using Apache Hadoop. This work explains the design of K-Means Algorithm using Mapper

* Asst. Prof., Department of Computer Application, Madras Christian College, Tambaram–600 059, *Email: mahima.jane@gmail.com*

** Asst. Prof., Department of Computer Science and Engineering, Bharathidasan University, Trichy-620 023, *Email: georgeprakashraj@yahoo.com*

and Reducer routines It also provides the steps involved in the implementation and execution of the K-Means Algorithm.

Dilpreet Singh et al.,[3]. This paper analyses different hardware platforms available for big data and assesses the advantages and disadvantages of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support. In addition a detail view on the software in different platformswere also studied.

N. Vishnupriya et al.[5], In this method K Means Algorithm is used to cluster the data for different type of datasets in Hadoop framework and calculate the Sum of Squared Error value for the given data.

Mugdha Jain et al. [4]. Describes modified algorithm based on k-means. It is an innovative method for big data analysis which can be very quick, scalable and highly accurate. It deals in overcoming the disadvantages of k-means of uncertain number of iterations by fixing the number of iterations, without losing the precision.

Xiao Cai, et al[9].propose a new robust multi-view K-means clustering method to integrate heterogeneous features for clustering. Gopi Gandhi, Rohit Srivastava [10] suggests different Partitioning techniques, such as kmeans, kmedoids and clarans. In terms of dataset attributes, the objects with in single clusters are of similar characteristics where the objects of different cluster have dissimilar characteristics

3. CLUSTERING TECHNIQUES IN BIG DATA

Big data clustering techniques has two major categories namely Single Machine Clustering Technique and Multi-Machine Clustering Technique [1].

In Fig.1 the list of clustering technique and the Algorithmsare shown.

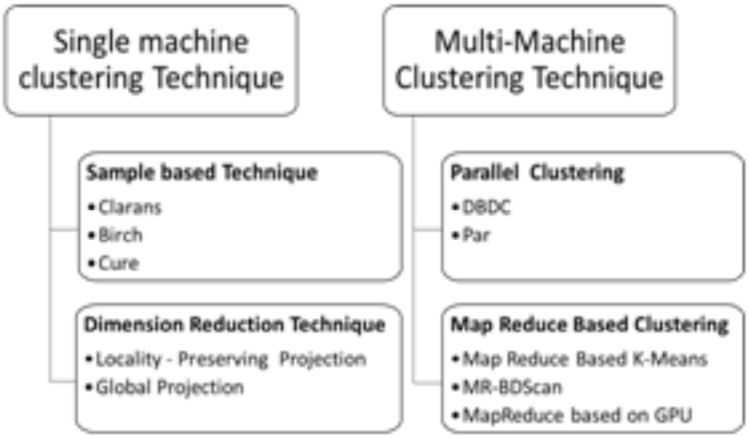
Single machine clustering Technique
Multi-Machine Clustering Technique
Sample based Technique
•Clarans
•Birch
•Cure
Parallel Clustering
•DBDC
•Par
Dimension Reduction Technique
•Locality - Preserving Projection
•Global Projection
Map Reduce Based Clustering
•Map Reduce Based K-Means
•MR-BDScan
•MapReduce based on GPU

Figure 1: Clustering Techniques

This paper considers the Multi-Machine clustering technique. The Multi-Machine clustering techniques are preferred to be most effective technique due to more scalability and better response time.

3.1. Hadoop Map Reduce

MapReduce [6] is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce [6] is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

3.2. Traditional K Means Algorithm

The basic idea of K-means algorithm is given by S. Yu, L. C. Tranchevent and X. H. Liu and T. W. Chen, C. H. Sun, H. H. Su, S. Y. Chien, D. Deguchi, I. Ide and H. Murase[7,8]. In K-Means algorithm, the objects are grouped based on the distance from the centre of each cluster and update it to the nearest cluster. The Average of each cluster is calculated and this process is repeated until the objective function converges. Here the objective function is the sum of the distance of each point to its centroid.

The Traditional K means algorithm is given is given below.

Step 1: Randomly select k data objects from data set D as initial centers.

Step 2: Repeat;

Step 3: Calculate the distance between each data object di $(1 \Leftarrow i \Leftarrow n)$ and all k clusters $C\,j(1 \Leftarrow j \Leftarrow k)$ and assign data object di to the nearest cluster.

Step 4: For each cluster j $(1 \Leftarrow j \Leftarrow k)$, recalculate the cluster center.

Step 5: Until no change in the center of clusters.

The disadvantages of the Traditional K-Means algorithm is that it works well for numeric data and not for other data sets. The initial clustering centroid is random

3.3. Enhanced K-Means Algorithm

The Enhanced K-Means Algorithm is an extension of the Traditional K-Means Algorithm. In the Enhanced K-Means clustering [11] algorithm, the centroids are fixed and the number of clusters are fixed. Hence it is easier than traditional K-Means clustering algorithm.

The Enhanced K-Means algorithm is given below in two phases.

Phase-I: To find the initial clusters

Step 1: Find the size of cluster Si $(1 = i = k)$ by Floor (n/k). Where n = number of data points Dp $(a1, a2, a3 \ldots an)$ K = number of clusters.

Step 2: Create K number of Arrays Ak

Step 3: Move data points (Dp) from Input Array to Ak until Si = Floor (n/k).

Step 4: Continue Step 3 until all Dp removed from input array

Step 5: Exit with having k initial clusters.

Phase-II: To find the final clusters

Step 1: Compute the Arithmetic Mean M of allinitial clusters Ci

Step 2: Set 1 j k

Step 3: Compute the distance D of all Dp to M of Initial Clusters Cj

Step 4: If D of Dp and M is less than or equal tootherdistances of Mi (1 i k) then Dp stay in same cluster Else Dp having less D is assigned to Corresponding Ci

Step 5: For each cluster Cj (1 j k), Recomputethe M andmove Dpuntil no change inclusters.

The disadvantage of Enhanced K-Means algorithm is it takes time as it is performed in single machine.

4. SBKMA: SORTING BASED K-MEANS CLUSTERING ALGORITHM USING MULTI MACHINE TECHNIQUE

The Traditional K Means does not use Sorting and uses only Single Machine Clustering Technique and the Enhanced K-Means algorithms uses Searching and not sorting. Also, it uses only the Single Machine Clustering Technique. These issues are considered and overcome in this proposed SBKMAmethod. 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. Hence number of iterations is considerably reduced. In addition to this enhancements, Multi-Machine clustering technique [1]allows 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. Multi Machine technique using Hadoop reduces iterations and decreases the execution time.

The Proposed SBKMA Algorithm is given below.

Step 1: Load the data set

Step 2: Choose k clusters

Step 3: Sort the data set

Step 4: Calculate the mean and chose centroids based on K number of cluster divided.

Step 5: Find the distance between the objects and centroids

Step 6: Group objects with minimum distance

Step 7: Repeat step 4, 5 and 6 till the clusters have no change

Step 8: Stop the program

The Proposed algorithm SBKMA: Sorting based K-Means Clustering Algorithm using Multi Machine techniqueis given as a flowchart in figure 2

5. EXPERIMENTATION AND ANALYSIS

The Traditional K-Means algorithm, Enhanced K-Means Algorithm and Proposed SBKMA algorithm are implemented in Hadoop MapReduceusing Multiple Machine Technique. Initially ajava file is created for random generation of 1 TBofmobile dataset in the form of text files. This data is consumed by the Hadoop Map Reduce and the three algorithms are compared for analysis.

Figure 3. gives the results on the time taken for execution by the three algorithms.

The results empirically concludes SBKMA algorithm consumes very less time for execution compared to other Algorithms.

Figure 4 gives the number of iterations compared along with other algorithms. It is very clear that the SBKMA algorithms has a very less iteration compared to the rest of the algorithms.

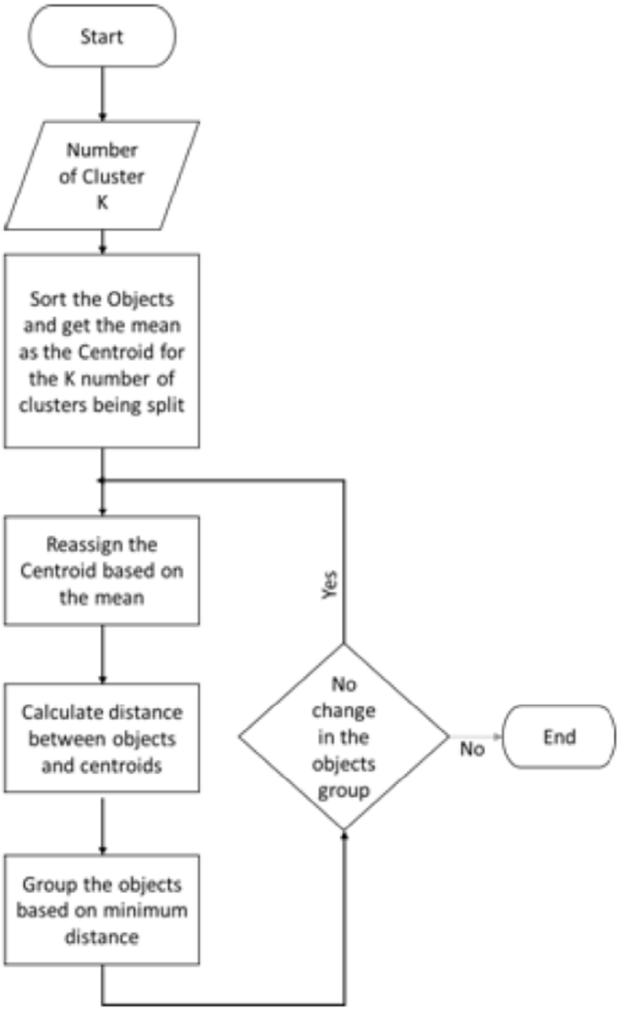
Start
Number of Cluster K
Sort the Objects and get the mean as the Centroid for the K number of clusters being split
Reassign the Centroid based on the mean
Calculate distance between objects and centroids
Group the objects based on minimum distance
No change in the objects group
Yes
No
End

Figure 2: Proposed SBKMA Algorithm

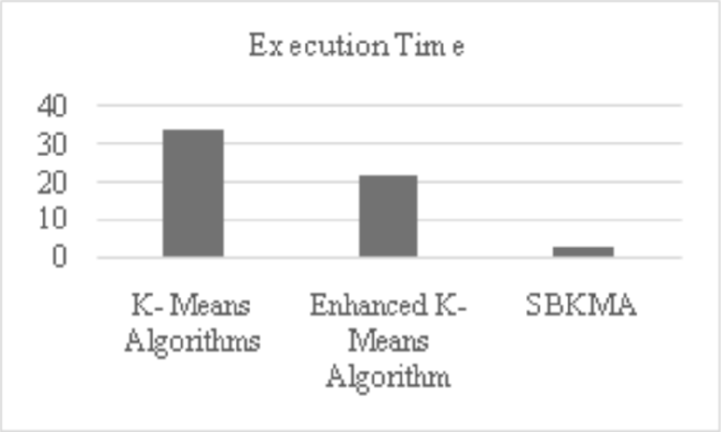

Figure 3: Execution Time (in secs)

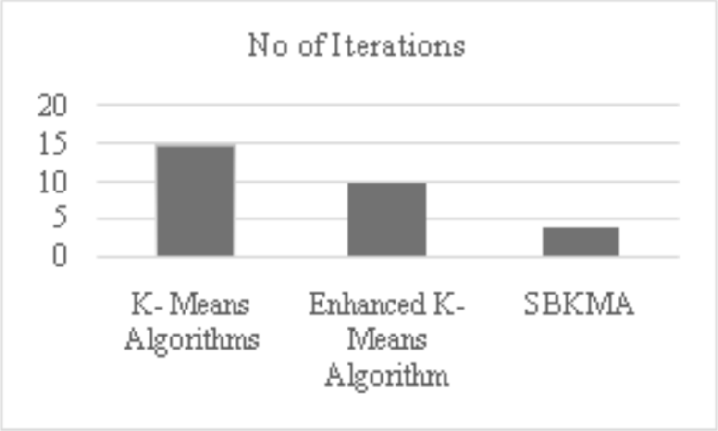

Figure 4: Number of Iterations

6. CONCLUSION

The Algorithm is discussed in the paper is tested with a dataset using Hadoop and MapReduce and is found that it decreases the execution time and the number of iterations compared to other existing algorithms. This can be further extended to get optimal K value with less number of clusters.

Dwindling down to conclude, the paper empirically proves that the proposed K-means based algorithm- SBKMA: Sorting based K-Means Clustering Algorithm using Multi Machine Technique overcomes the inefficiencies in the traditional methods.

References

[1] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah and TututHerawan, “Big Data Clustering: A Review”, Research Gate, Jun, (2014).

[2] Yaminee S. Patil, M. B. Vaidya “K-means Clustering with MapReduce Technique”, International Journal of Advanced Research in Computer and Communication Engineering, Nov, (2015).

[3] Dilpreet Singh, Chandan K Reddy, “A survey on platforms for big data analytics”, Journal of Big Data, (2014).

[4] Mugdha Jain, ChakradharVerma, “Adapting k-means for Clustering in Big Data”, International Journal of Computer Applications, Sep, (2014).

[5] N.Vishnupriya, Dr. F. Sagayaraj Francis, “Data Clustering using MapReduce for Multidimensional Datasets”, International Advanced Research Journal in Science, Engineering and Technology, Aug, (2015**)**.

[6] http://www.tutorialspoint.com/hadoop/

[7] S. Yu, L. C. Tranchevent and X. H. Liu, “Optimized Data Fusion for Kernel K-means Clustering”, Pattern Analysis and Machine Intelligence, (2012), pp. 1031 – 1039.

[8] T. W. Chen, C. H. Sun, H. H. Su, S. Y. Chien, D. Deguchi, I. Ide and H. Murase, “Power-Efficient Hardware Architecture of K-means Clustering With Bayesian Information Criterion Processor for Multimedia Processing Applications”, Emerging and Selected Topics in Circuits and Systems, (2011), pp. 357–368.

[9] Xiao Cai, FeipingNie, Heng Huang, “Multi-View K-Means Clustering on Big Data” Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.

[10] Gopi Gandhi, Rohit Srivastava, “Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms”-International Journal of Computer Applications (0975–8887) Volume 87–No. 9, February 2014.

[11] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity” Middle-East Journal of Scientific Research 12 (7): 959-963, 2012.