FPBCA: Framework for Partition Based Clustering Algorithms

[1]E. Mahima Jane,[2]Dr. E. George Dharma Prakash Raj[2]

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

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

Abstract-The term Big Data is practical to data sets whose sizes are beyond the ability of traditional relational data bases to capture, manage, and process. The following are the characteristics of big data– high volume, high velocity, or high variety. Big data comes from a variety of sources such as networks, transactional applications, web, and social media - much of it generated in real time and in a very large scale. Clustering is technique to examine large data. This paper proposes a novel framework for the existing partition based clustering algorithms by reducing the iterations and execution time.

Keywords- *Clustering, Big Data;Partition.*

I. INTRODUCTION

Clustering is the task of separating the 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. A Cluster is a set of entities which are alike, and entities from different clusters are not alike. The partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster. These clusters should fulfill the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group.[1]This paper is structured as follows Section II discusses about the related work, section III is about the sorting based partition based algorithms, Section IV discusses the framework and Section V concludes the paper.

II. RELATED WORK

Yugal Kumar and G. Sahoo *[1]*, suggested a method to deal with the initial centroid problem in K-Means algorithm based on binary search. Binary search is one of the searching methods that is used to find an item in given list of data items. Arash Ghorbannia Delavar*et al.* [2], proposed an algorithm to find the initial cluster centers by selecting two attributes which describes the data space better using the number of neighbors in a specific radius. Mahima Jane et al[4] have performed a survey on various partition based clustering algorithms. Shah et al[7]., proposed a model for distributed document clustering to address the scalability issue. Kavya *et al*[8] In this paper Performance is evaluated in both sequential and parallell considering different iterations and calculating the elapsed time in the iterations process.

III. PROPOSED PARTITION BASED CLUSTERING ALGORITHMS

SBKMA: Sorting based K- Means Clustering Algorithm using Multi Machine Technique for Big Data[3] is an efficient partition based clustering algorithm which minimizes the execution time and reduces the number of iterations by fixing the initial centroid values. SBKMA algorithm loads the data into the number of nodes given. Each partition are sorted by the attributes given. All the partitions are merged to form sorted dataset. The number of cluster K is randomly generated. Depending on the size of the K the sorted data are partitioned into equal size. Mean of each partition is calculated and taken as initial centroids. Distance is calculated using Euclidean distance. Objects are compared with the initial centroids. Objects are grouped with the nearest cluster. Distance calculation and mean calculation for centroids are repeated till there is no change in the cluster formation.

```
SBKMA Algorithm
Step 1:Start
Step 2:Load the dataset into the multiple nodes
Step 3: Generate random value for clusters K
Step4: Divide the dataset D into number of nodes n
        Each node is sorted with the pivot element
Step 5: Sorted data Si are divided into K Random generated
Step 6: Mean Mi of every partition is calculated
Step 7: Mean of the datapoints dp is taken as centroids of each cluster
Step 8: Compute the distance between each data point di (1<= i<= n) to all the initial centroids cj (1 <= j <= k).
Step 9: For each data point di, find the nearest centroid cj and assign di to cluster j.
Step 10: Set ClusterNo[i]=j.
Step 11: Set Clustergroup[i]= d(di, cj).
Step 12: For each data point di,
        Compute the distance from the centroid to the nearest cluster
    If this distance is less than or equal to the presentcentroid the data point stays in the same cluster.
```

```
  Else
  Compute  the distance d(di, cj) and recalculate the centroid .
  End for;
Step 13: Repeat step 9 to 12 till there is no change in the cluster formation.
Step 14: End
```

SBKMEDA: Sorting based K- Median Clustering Algorithm using Multi Machine Technique for Big Data[4] is an efficient partition based clustering algorithm which reduces the execution time even when the data are skewed.SBKMEDA algorithm loads the data into the number of nodes given. Each partition are sorted by the attributes given. All the partitions are merged to form sorted dataset. The number of cluster K is randomly generated. Depending on the size of the K the sorted data are partitioned into equal size. Median of each partition is calculated and taken as initial centroids. Distance is calculated using Euclidean distance. Objects are compared with the initial centroids. Objects are grouped with the nearest cluster. Distance calculation and mean calculation for centroids are repeated till there is no change in the cluster formation.

SBKMEDA Algorithm

```
Step 1:Start
Step 2:Load the dataset into the multiple nodes
Step 3: Generate random value for clusters K
Step4: Divide the dataset D into number of nodes n
          Each node is sorted with the pivot element
Step 5: Sorted data Si are divided into K Random generated
Step 6: Median Mi of every partition is calculated
Step 7:  Median of the datapoints dp is taken as centroids of each cluster
Step 8: Compute the distance between each data point di (1<= i<= n) to all the initial centroids cj (1 <= j <= k).
Step 9:  For each data point  di,  find the  nearest  centroid cj and assign di to cluster j.
Step 10: Set ClusterNo[i]=j.
Step 11: Set Clustergroup[i]= d(di, cj).
Step 12: For each data point di,
            Compute the distance from the centroid to the nearest cluster
  If this distance is less than or equal to the present centroid the data point stays in the same cluster.
  Else
Compute  the distance d(di, cj) and recalculate the centroid .
End for;
Step 13: Repeat step 9 to 12 till there is no change in the cluster formation.
Step 14: End
```

SBKMMA : Sorting based K Means and Median based Clustering Algorithm using Multi MachineTechnique for Big Data[5] is an efficient clustering based algorithm where the execution time decreases for whatever data is loaded and reduces the iterations by initializing the centroid values. This algorithm loads the data into the nunber of nodes given. Each partition are sorted by the attributes given. All the partitions are merged to form sorted dataset. The number of cluster K is randomly generated. Depending on the size of the K the sorted data are partitioned into equal size. Mean and Median of each partition is calculated. When the difference between the mean and median are more median will be taken as centroids else the value of mean will be taken as centroids. Centroids are initialsed to the objects which they belong. Distance is calculated using Euclidean distance. Objects are compared with the initial centroids. Objects are grouped with the nearest cluster. Distance calculation and mean calculation for centroids are repeated till there is no change in the cluster formation.

SBKMMA : Sorting based K Means and Median based Clustering Algorithm using Multi Machine Technique for Big Data

```
Step 1:Start
Step 2:Load the dataset into the multiple nodes
Step 3: Generate random value for clusters K
Step4: Divide the dataset D into number of nodes n
          Each node is sorted with the pivot element
Step 5: Sorted data Si are divided into K Random generated
Step 6: Mean and Median of every partition is calculated.
Step 7: If the value of mean and median differs more
            Median will be taken as initial centroids
      Else
      Mean will be taken as initial centroids
Step 8: Initial data points are assigned to the centroids cj of the clusters they belong.
Step 10: Set ClusterNo[i]=j.
Step 11: Set Clustergroup[i]= d(di, cj).
Step 13: For each data point di,
Compute the distance from the centroid to the nearest cluster
If this distance is less than or equal to the present centroid the data point stays in the same cluster.
Else
Compute  the distance d(di, cj) and recalculate the centroid .
End for;
Step 14: Calculate the distance between the objects to all the centroids and assign it to the nearest
Step 15: Repeat step 10 to 14 till there is no change in the cluster formation.
Step 15: End
```

IV. FRAMEWORK OF SORTING BASED PARTITION BASED CLUSTERING ALGORITHMS

A Framework for Efficient Partition based Clustering Algorithms is developed to integrate the proposed partition based Clustering Algorithms to enhance the performance of Big Data Applications. The proposed algorithms are taken into consideration to design this framework for reducing the execution time and increase the speed. This framework is systematized based on the different types of large datasets. The proposed algorithms SBKMA, SBKMEDA and SBKMMA are used to provide a better execution time.

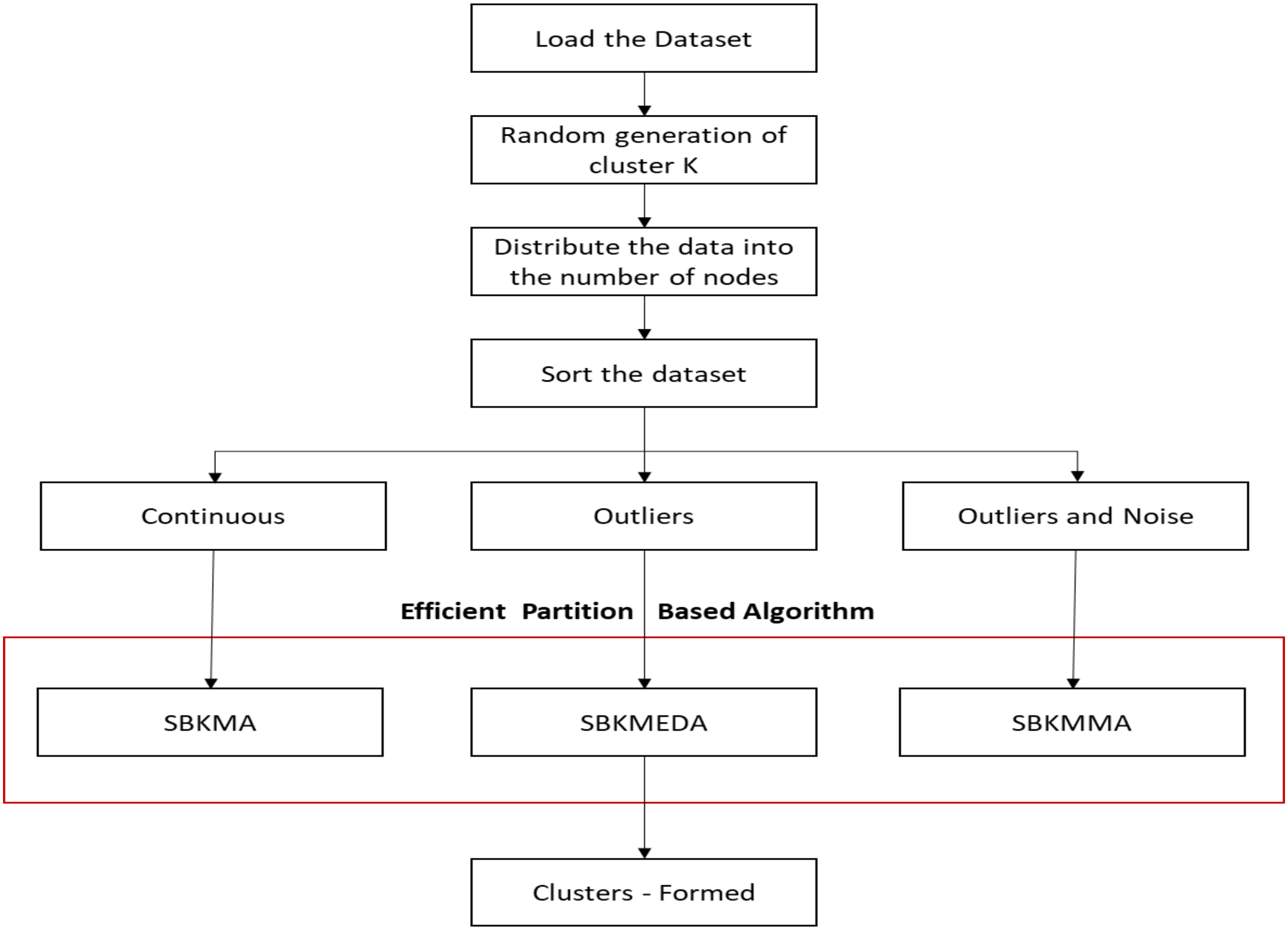
Load the Dataset
Random generation of cluster K
Distribute the data into the number of nodes
Sort the dataset
Continuous
Outliers
Outliers and Noise
Efficient Partition Based Algorithm
SBKMA
SBKMEDA
SBKMMA
Clusters - Formed

Fig.1:Framework for Partition Based Clustering Algorithms

The proposed framework is given in Figure1.First part loads the dataset. The data are partitioned into the number of nodes given. Sorting is performed in all the nodes. Based on the data when it is continuous mean is taken as initial centroids in the SBKMA algorithm. When the sorted data is skewed median is taken as initial centroids in the SBKMEDA algorithm. Initial centroids allow the algorithm to cluster the data quickly. Mixed data uses mean and median value to fix the centroids. Depending on the value mean or median is taken for the various partitions. These values are assigned as the initial centroids to the objects in the SBKMEDA algorithm. Final efficient clusters are formed with reduced iteration and execution time.

V. CONCLUSION

In this paper a Framework for Partition based ClusteringAlgorithms is proposed for Big Data. The ultimate aim of those algorithms is to reduce they execution time by doing it distributed and sorting solves the drawback of iterating data points using sorting. The results of these algorithms are confined and give better result. In our future work, it is proposed to improve our framework by considering other factors of fixing the clusters and to support other categorical data also.

REFERENCES

[1]. Yugal Kumar and G. Sahoo " A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science andTechnology**,** Vol. 9(2014),pp.43-54

[2]. Arash Ghorbannia Delavar and Gholam Hasan Mohebpour, "ANR: An algorithm to recommend initial cluster centers for k-means algorithm", Journal of mathematics and computer science 11 (2014), 277- 290.

[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 " in the Advances in Intelligent Systems and Computing, vol 645. Springer,April 2018.

[5]. Mahima Jane and Dr. E. George Dharma Prakash RajSBKMMA: Sorting Based K Means and Median Based Clustering Algorithm Using Multi Machine Technique for Big Data. International Journal of Computer (IJC), [S.l.], v. 28, n. 1, p. 1-7, jan. 2018. ISSN 2307-4523.

[6]. E. Mahima Jane and E. George Dharma Prakash Raj,"Survey on Partition based Clustering Algorithms in Big Data", International Journal of Computer Sciences and Engineering, Vol.5, Issue.12, pp.323-325, 2017.

[7]. Neepa Shah and Dr. Sunita Mahajan "Distributed Document Clustering Using K-Means" International Journal of Advanced Research in Computer Science and Software Engineering 4(11), November - 2014, pp. 24-29

[8]. Kavya D S 1, Chaitra D Desai2 "Comparative Analysis of K means Clustering Sequentially And parallely "International Research Journal Of Engineering And Technology (IRJET) E-ISSN: 2395 -0056VOLUME: 03 ISSUE: 04 | APR-2016.