ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Modified K-Means Algorithm on Hadoop

Sreedevi KM[1], Mr.Hareesh MJ[2], Honeytta Kunjachan [3]
Computer Science Department, KTU University Kerala, India
sreedevikunnathmana@gmail.com
brijithoney@gmail.com
hareeshmjoseph@gmail.com

ABSTRACT

Now datas are rapidly increases day by day. The data storing is very difficult. To overcome this problem using BigData Technology. The datas are saved in cluster form. The clustering is grouping the similar objects. The clustering algorithm is used for different purpose. The traditional K-Means algorithm, randomly select the initial centres and their result is not accurate. The proposed algorithm as advanced K-Means algorithm. In advanced K-Means algorithm, initial centres takes based on dimensional density. The implementation done by 2 node clustering on hadoop MapReduce function. The MapReduce, first take Map function and then Reduce function. MapReduce job take less time.

Keywords—Hadoop, Big Data, Clustering, K-Means clustering algorithm, Data point.

I. INTRODUCTION

Clustering is grouping the similar objects. Different methods used in clustering algorithm. K-Means is mostly used in clustering algorithm. Data mining is major role in clustering. Data mining is mine the data from dataset and result is meaningful. The clustering applied to fraud detection, market analysis etc.

The K-Means clustering algorithm, randomly select the initial centroids [2]. Based on these centroids find the cluster. The number of clusters based on the value of K. The input of K-Means algorithm, K and Dataset. The K represent number of clusters. The output is number of clusters. The advantage of traditional K-Means algorithm as simplicity and speed. The disadvantage of this algorithm gives different outputs of each run because of its random initialization problem.

The proposed K-Means algorithm based on dimension density [1]. This algorithm overcome the problem of traditional K-Means algorithm. The initial datapoints selects from highly dense area. Therefore, the result is stable. This paper discuss modified K-Means algorithm with intro- duction. The section 2, deals with various clustering methods. Section 3, background, it include big data, hadoop, K-Means clustering.Section 4,deals proposed method.Section 5 discuss experiment implementation and results. Section 5, concludes the paper.

II. CLUSTERING METHODS

The main idea behind clustering is to maximize the intra cluster similarity and minimize the inter cluster similarities [2]. The clustering is unsupervised learning, which do not have predefined classes. The types of clustering are distribution based methods, centroid based methods, connectivity based methods and density models.

- Distribution based methods
  Expectation-Maximization method
- Centroid based method
  K-Means algorithm
- Connectivity based methods
  Hierarchical algorithm
- Density Model
  DBSCAN

III. BACKGROUND

In this section, discuss about Big data, Hadoop, MapReduce and K-Means clustering algorithm.

A. Big Data

Big data contain large volume of data. The property of big data describes Vs. They are Velocity,Volume,Variety. Velocity means speed of data generation and processing. Big data implies volume of data. In Internet technology, every seconds data is generated. In the variety different types of data is used. The variety of datas are pictures, audios, videos etc.

There are two more dimensions that characterizes the big data: Variability and Complexity [5]. Variability means data flows are not persistent. Complexity means, in big data using various data sources. The main characteristics of Big Data as

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

speed, Scalability, Security etc. [6].

B. Hadoop

Apache Hadoop is developed to scale up from single servers to a clusters of multiple machines, each of these offering its own(local) computation and storage capabilities and Apache Hadoop is a Java based open source software [1]. Doug Cutting and Mike Cafarella have created the Hadoop framework in 2005 [7]. Hadoop is a composition of following two components are HDFS (Hadoop Distributed File System) and MapReduce [8].

Subject	A	B
1	1	1
2	2	1
3	4	3
4	5	4

HDFS allow to store efficiently numerous large files across thousands of machines and access every data chunk in parallel manner [8].

C. MapReduce

It is a framework of Hadoop. Hadoop provides multiple nodes of cluster. MapReduce has two stages

1) Map function
2) Reducer function

After completion of map phase, then do Reduce phase. The dataset is partitioned into various parts. The Map function applied this different parts. Then sorts the output. The reduce function applied to these results. After the completion of reduce function, the output is stored in a le system. The input and output of MapReduce function as *key, value* pair format.

D. K-Means Clustering Algorithm

The K-Means algorithm is a partitioned method or centroid method. The term ”K-Means”, first used by James MacQueen and standard algorithm is first proposed by Stuart Lloyd [1].

The K-Means algorithm follows as [1]

1) Choose the number of cluster value K.
2) Randomly select K initial point centroids.
3) Using Euclidean distance, compute the distance between centroid and data object.

$$Euclideandistance = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \quad (1)$$

4) Recalculate the centroid in each cluster.
5) Repeat the step until centroids are equal.

The K-Means explained with example. Given a dataset consisting of the scores of two individuals. Their corresponding table is shown in below Table 1. The individual A and B got marks for subject 1,2,3,4.

First, choose the value of K and randomly select initial

centroids. In this example, initial centroids are C1 = (1,1) and C2 = (2,1) where the C1 and C2 are initial centroids. Then, find the centroid distance of each data point using Euclidean distance. For example, subject 3’s centoid distance are C1 = 3.61 and C2 = 2.83. And calculate other datapoints. After the first iteration, cluster 1 includes subject 1 and cluster 2 includes subject 2,3,4. Calculate new centroid based on the clusters. New cluster centroids are C1 = 1,1) and C2 = (3.6, 2.6). And compute new cluster datapoints. After the completion of second iteration, the clusters are not converge. So cluster 1 contain Subject 1 and 2 and cluster 2 contain Subject 3 and 4.

1) MapReduce K-Means clustering: The map function determines which are the closer to the centroids. The reducer function, is output of mapper function and compute new K centroids. The output of K Means as ¡key, value¿ format.

IV. PROPOSED METHOD

The traditional K-Means algorithm select initial centroids as randomly. In random initial centroid, each time we get different cluster results. So the cluster is unstable. The proposed modified K-Means clustering algorithm, select the initial centroid based on dimensional density. The basic idea behind this modified K-Means clustering algorithm is that we select optimal K datapoints in the highly dense areas as the initial center for the algorithm [1]. Therefore the

algorithm provides stable cluster. The advantage of this method, unwanted cluster removed in final cluster. To execute the algorithm on same dataset, we get correct outputs. The flow diagram of modified algorithm is given below [Fig.1]. In High density area, select initial cluster centroids. Then, compare datapoint in cluster center and distance between datapoint. And get K clusters.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

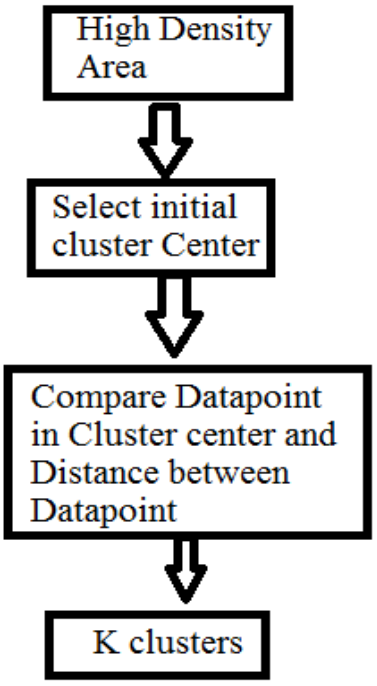

Fig. 1. Flow diagram of Modified K-Means clustering algorithm

A. Algorithm
The implementation of modified K-Means algorithm,
first take Map function and then do Reducer function. The algorithm is explained below [1].
Input : Dataset, K, Threshold value
Output : K cluster
Mapper function

1)Compute the distance between each datapoint to all other points in the dataset.

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 ....} \qquad (2)$$

where dij is distance between two data point i and j and xi2 means the distance of second dimension xi.
2) Calculate average value R

$$R = \frac{\sum i \leq ni < j \leq n(d(ij)}{n * \frac{(n-1)}{2}} \qquad (3)$$

3) Calculate threshold value.
4) Check data points which belongs high density area.
a) If yes, assign to high density area.
b) Else, assign low density area.
Reducer function

1) Find K farthest data points from high density area
a) These K farthest data points are sorted by decreasing order.
b) Select two data points which is initial cluster center set.
Selection of two data points based on largest distance in high density area.
c) Compare data points in initial cluster center and distance between data points in high density area.
d) Repeat this steps until reach K clusters.

V. RESULTS
A. Dataset
In this implementation using ”20 newsgroup” dataset. This dataset consist of 20000 messages taken from 20 newsgroup [4]. The folder realated as autos, football, electronics etc. are the various topics. In each folder contain other files. This dataset contain attributes like from, newsgroup, messageID etc. The dataset available from UCI machine learning Dataset repository.

B. Experimental Setup
To implement this algorithm we need
1) 1 Master node PC (Ubuntu 14.04 LTS)
2) 1 Slave node PC (Ubuntu 14.04 LTS)
3) Hadoop 2.7.1
4) Eclipse 4.3.0
5) JDK 1.8
The modified K-Means algorithm implemented on 2 node clustering. It include 2 networks. One PC is act as Master node and other node working as slave node. The advantage of single node clustering to two node clustering, running take less time. In two node clustering, Master node splitts the jobs equally shared to slaves. The Namenode, which running on master node. The Datanode running on Slavenode. First set up the hadoop environment and start all services using ”./start-all.sh” command.
C. Screenshots
The Fig.2 shows key value pair of each data point. For
example, 30 is the key and their corresponding value is
0.002344. The key represent document and value represent value of that document.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Fig. 2. Key Value pair of datapoint

The Fig.3 shows K clusters. In this figure the K value is 10. Based on the cluster centroid, 10 clusters are formed.

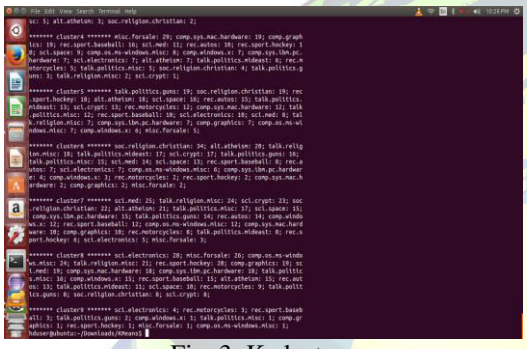

Fig. 3. K clusters

VI. CONCLUSION

The traditional K-Means clustering algorithm, randomly select the cluster centroids. The random selection which tends to unstable results. It gives empty clusters, even if any datapoint belongs to that cluster. In the cluster, if the outliers are present the centroids may not be representative. Difficult analyse whether datapoints belongs to the particular cluster. In modified K-Means clustering based on dimension density. It contain high density area and low density area. This algorithm, only considered about high density areas. In high density area, calculate the two cluster centroids, which is farthest points. Then compute distance centroids and each datapoint. The advantage of Modified K-Means, remove the unwanted datapoints in the cluster.

VII. FUTURE WORK

It implemented on three node and 4 node clustering. And compare the results. To modify the distance computation.

REFERENCES

[1] Nadeem Akthar, Mohd Vasim Ahamad and Shahbaz Khan ' Clustering on Big Data Using Hadoop MapReduce ", IEEE 2015.

[2] Prajesh P Anchalia, Anjan K Koundinya, Srinath N K, " MapReduce Design of K-Means Clustering Algorithm ", IEEE 2013.

[3] Botcha Chandrasekhara Rao , Medara Rambabu " Implementing KMeans Clustering Algorithm Using MapReduce Paradigm ", International Journal of Science and Research 2015. .

[4]https://archive.ics.uci.edu/ml/datasets/Twenty +Newsgroups

[5] A, Katal, Wazid M, and Goudar R.H. " Big data: Issues, challenges, tools and Good practices ", Noida: 2013, pp. 404 409, 8-10 Aug. 2013.

[6] Venkata Narasimha Inukollu , Sailaja Arsi, Srinivasa Rao Ravuri " Security Issues Associated With Big Data In Cloud Computing ", International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[7] " Apache Hadoop " http://hadoop.apache.org/

[8] Borthakur, D. " The Hadoop Distributed File System: Architecture and Design ", 2007.