International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 6, December 2016, pp. 3047~3051
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i6.11207

Issues of K Means Clustering While Migrating to Map Reduce Paradigm with Big Data: A Survey

K. R. Nirmal, K. V. V. Satyanarayana
Koneru Lakshmaiah Education Foundation (KLU), India

Article Info

Article history:

Received May 9, 2016
Revised Sep 20, 2016
Accepted Oct 4, 2016

Keyword:

Big data
Categorial data set
Clustering
Data mining
Initial centroid selection
K means clustering
Map reduce paradigm
Parameter k in k-means

ABSTRACT

In recent times Big Data Analysis are imminent as essential area in the field of Computer Science. Taking out of significant information from Big Data by separating the data in to distinct group is crucial task and it is beyond the scope of commonly used personal machine. It is necessary to adopt the distributed environment similar to map reduce paradigm and migrate the data mining algorithm using it. In Data Mining the partition based K Means Clustering is one of the broadly used algorithms for grouping data according to the degree of similarities between data. It requires the number of K and initial centroid of cluster as input. By surveying the parameters preferred by algorithm or opted by user influence the functionality of Algorithm. It is the necessity to migrate the K means Clustering on MapReduce and predicts the value of k using machine learning approach. For selecting the initial cluster the efficient method is to be devised and united with it. This paper is comprised the survey of several methods for predicting the value of K in K means Clustering and also contains the survey of different methodologies to find out initial center of the cluster. Along with initial value of k and initial centroid selection the objective of proposed work is to compact with analysis of categorical data.

Corresponding Author:

K. R. Nirmal,
Department of Computer Science & Engineering,
KL University,
Vaddeswaram-522502, Guntur Dist., A.P, India.
Email: khyatinirmal@gmail.com

1. INTRODUCTION

Big data is a term indicates the rise in the volume of data that are challenging to store, process, and analyze through traditional database technologies. Big Data is generated on a notable rate on daily basis. An International Data Corporation (IDC) report predicts that ‘‘from 2005 to 2020, the digital universe will grow by a factor of 300, from 130 Exabyte to 40 000 Exabyte’’ and that ‘‘from now until 2020 will about double every two years” [1]. The form of Data collected from different sources like media, sensor nodes, banking transaction, IoT based devices etc. are having major dissimilarities. Xindong Wu et. al has propose a HACE theorem to model Big Data characteristics. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data [2]. Ordinarily software tools are not sufficient enough to store, capture and analyze the data.

These will extent the new direction in research. In any computerized application the safe decision making is based on this collected data so only storing and maintaining of data is not adequate. There is the necessity to identify the significant data with proper semantic meaning, finding out the pattern from data or separating the data in groups.

“Extracting or mining" knowledge from large amounts of data is defined as Data Mining [3]. To split up the data according to their similarities based on multiple characteristic is known as clustering. The

well-defined clustering will depend on the algorithm selected for the grouping. K Means Clustering is the most commonly used algorithm but the incorrect parameters for initial values will degrade the performance. This paper address the issues related to K Means Clustering by relative survey. Essential Technologies for migrating K Means on map reduce are described in Section 2, Furthermore in Section 3 it contains the comparative study of different methodologies used for migrating K means Clustering on Map Reduce paradigm and we have identified issues related to K Means Clustering and Section 4 contains our proposed objective outline to address all the issues related with K Means Clustering and Section 5 is conclusion of our work.

2. ESSENTIAL TECHNOLOGIES

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [3]. Dissimilarities between objects can be calculated by the variety of attributes associated with the objects. In general clustering algorithm is categorized into Partition based, Hierarchical method, Density based methods, Grid based methods, Model based methods etc. In partition based clustering k means clustering and its variations are most widely used.

2.1. K Means Clustering algorithm

The core objective of K Means clustering algorithm is to divide given n number of data objects into k number of cluster such that intra cluster similarity is high but the inter cluster similarity is low. The detailed algorithm is as below:

Algorithm:

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

k : the number of clusters,
D : a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar based on the mean value of the objects in the cluster;.
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;[3]

This algorithm is reasonably proficient for large datasets because time complexity is O(nkt), where n is number of objects in data set, k is the number of cluster and t is the number of iterations. Normally, $k \leq n$ and $t \leq n$.

2.2. MapReduce Pardigm

The parallel programming on single machine is not efficient for processing the Big Data. For Parallel programming frame work on multiple machines MapReduce is an essential preference. MapReduce is a programming model for Big data analyzing and processing. The core idea is to divide bulky tasks into many small tasks and conquer them after processed. Map and Reduce are two phases of the programming paradigm. The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation as two functions: map and reduce.

Map Function is composed by the user and accepts input as a <key, value> pair and produces a set of intermediate <key, value> pairs. From all the intermediate <key, value> pairs values associated with the same intermediate key are combined <key, list of value> and passes to the reduce function. The reduce function, also composed by the user, accepts an intermediate result <key, list of value>. It merges these values together to form a conceivably reduced set of values. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory [4].

3. RELATED WORK

Big data analysis needs merging of techniques for Big Data and Machine Learning. K Means Clustering is the one of such algorithm offerings in together the fields. Many research studies have already done on this area , this section surveyed on it and mostly focus on three issues related to K Means Clustering while migrating on Map Reduce with Big data – Selection of Initial Value of K, Initial Centroid Selection, Managing the categorical data.

1. Yujjexu et.al [5] has proposed Efficient k-means ++ Approximation with MapReduce, choose the initial to achieve a solution that is provably close to optimal one. Here only one MapReduce job is used to initialize k centers , that avoids multiple round on MapReduce job on many machine. Here the initial value of k ist put on as prerequisite of algorithm , so improper value of k will may affect the complexity of algorithm.
2. Kehe Wu et.al [6] have done research and improve K means algorithm based on Hadoop and come out with the solution to define initial centroid. Using Convex hull and opposite Chung points the initial two cluster centers are defined. The optimal number of cluster k is decided using distance cost function. The text files are used to experiment the algorithm, so in real time application the new strategy has to be develop that will work for heterogeneous data set.
3. Anupama Chadha and Suresh Kumar [7] has presented An improved K- means Clustering Algorithm: A step forward for removal of dependency on K. The modified K –means algorithm does not require number of cluster (k) as input. Initially it selects the two centroids which are farthest apart, and considered as two initial centroids. The value of k is decided by using outliers which are originated during the calculation of Euclidean distance between every tuple and new means of the cluster centers. Accuracy accomplished by this algorithm is better than the original K-means algorithm. The downside of this algorithm is that it is not designed for map reduce paradigm and it works only for numeric datasets.
4. Prajesh Anchalia [8] has improved k-means clustering algorithm by introducing combiner. The combiner reads the output produced by mapper and calculates the centroid for each mapper. Now the reducer calculate the global centroid using the value of local centroid read from each mapper. The performance of Hadoop can be increased by cutting down the read and write operation from mapper and reducer respectively. Here Sartup algorithm is used to calculate the initial set of centroid which again requires the number of cluster (k) as input.
5. Arshad M. Mehar et.al [9] has introduced k means clustering algorithm for determining optimal value of k.Mostly distance based method are used to calculate the value of k. In contrast with other methodologies in this algorithm the joint probability is used. The fundamental idea is to analyze the movement of objects between clusters. The diagonally dominant probability matrix is produced using the movement of membership will be used to decide the set of range of optimal value for k clusters. Here the algorithm focus on synthetic data set for two dimensions and it is not proposed for Map Reduce programming.
6. Jing Zhang et. al [10] has designed a 2 tier clustering algorithm with Map-Reduce for distributed clustering environment. The entire procedure is distributed in four parts. Split phase divides the input file in m parts, where m is user defined. The one split clustering using k means algorithm is performed in map phase for each mapper and intermediate result is generated. The intermediate result are again partitioned into r region in Partition phase and assigned to reducer. Final result is calculated in reducer phase by using integration strategy of two tier clustering. Here also the size of r is defined by user.
7. Mohammad Kakooei and Hadi S.Shahhoseini [11] have presented a parallel k-means clustering initial center selection and dynamic center correction on GPU. In this algorithm initially group of initial centers are selected which should be less than the parallel stream execution provided by machine. It calculates the inner distance between data points and center of each group. The group having minimum distance is selected as initial center. For the main clustering the algorithm applies two asynchronous streams. The inner distance of main clustering is calculated by first stream and other stream calculates the inner distance of new initial centers. Then these two distances are compared and if newly generated distance is less than main distance then main cluster and newly generated clusters are combined into one group. The stream of main clustering procedure goes on continuously. In this algorithm initial the value of k is selected arbitrary, the improper selection for value of k will may be degrade the performance.
8. Poonam Ghuli et. al [12] has done a comprehensive survey on centroid selection strategies for distributed k- means clustering algorithm. The execution is divided into four modules. In Sorting Module the data set is processed and simplify it for selecting the initial centroid. Here for initialization of centroid three different strategies are proposed. 1) Weighted Average Sorting module calculate the scores using uniformly assigned weights to attributes of dataset. The sorting of data points is done in reducer according to the average value and write result into intermediate file. 2) Heuristic Sorting Module, In this module the attribute having highest rage is selected and reducer sorts the data points in

increasing order and writes results in intermediate file. 3) Principal Component Analysis, In this value the attribute having highest variance is selected and the reducer sorts the data points in increasing order of variance and writes the results in intermediate file. Initial centroid selection module splits the data set into k subsets. Now median of the data set is calculated and that will act as an initial centroid. This will act as input file to Map Reduce Algorithm. Iterative Clustering and cluster assignment module will act in regular manner to divide and categorize the dataset into k clusters. All over again In second module the k is not automated and still it is user depended.

9. Mugdha Jain and Chakradhar Verma [13] has adopted k-means for clustering in Big Data. Here dataset is represented using matrix where data points are represented by rows and attributes of data points represented by columns. In the situation where some attribute value for data point is missing, matrix will not define any values. Data points are arranged according to decreasing order of priority of dimensions and variance of dimensions is calculated. Assignment of data set according to primary dimension is followed by all secondary dimensions and outliers are calculated. Distance between centroid and outlier is used to categorize the data set into clusters. The number of cluster k is predefined it is not the part of algorithm and the priorities are decided by user instead machine learning approach can be integrated to decide priorities.
10. Yiu-Ming Cheung [14] has proposed k*-means: A new generalized k-means clustering algorithm. This algorithm achieves improved clustering without predetermining precise cluster number in two steps. By assigning seed point to every cluster and using learning rule with panelized mechanism the input is categorized in particular cluster. The shortcoming of the algorithm is that it is not designed for Map Reduce paradigm.

4. PROPOSED OBJECTIVE

The proposed objective of this paper is to apply clustering on Big Data set using Map Reduce Paradigm. While considering above mentioned issues, outline is planned. The below Figure 1 displays outline of the proposed objective.

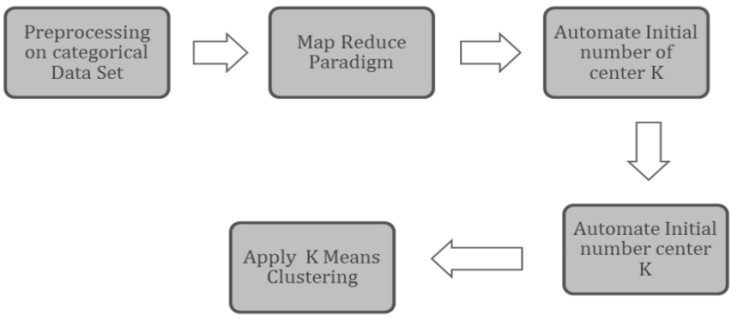

Figure 1. Outline of the proposed Objective

Preprocessing: At all times the Big data is always in the form of categorical data. Standard k means will may degrade in performance when considered for multidimensional data so Pre Processing on categorical data set is taken into the proposed objective.

Map Reduce Paradigm: To handle the Big Data Map Reduce paradigm is adopted to improve in performance and to improve scalability of data.

K selection module: The time complexity of k means algorithm is depended on predicated number of cluster, sometimes improper selection of value of k will degrade the performance so it is essential to automate it.

Centroid Selection Module: Another issue while adopting k means is to select initial centroid, which will huddle to adopt k means clustering. In proposed objective the initial centroid will be decided in algorithm only.

5. CONCLUSION

K Means Clustering is efficient algorithm to categorize the Big Data in to appropriate clusters. K Means cluster has some issues like initial number of cluster, initial selection of centroids. This article presents some issues to solve these issues which are helpful to predefine the declared parameters. Further this study also addresses the issues to adopt k means clustering for categorical datasets and based on this study one outline of objectives is proposed. Proposed model will apply selected strategies to predefine the parameters while handling the categorical data set.

REFERENCES

[1] Gantz J. and Reinsel D., "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, pp. 1-6, 2012.

[2] Wu X., *et al.*, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol/issue: 26(1), pp. 97-107, 2014.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2nd Edition, published by Elsevier.

[4] Dean J. and Ghemawat S., "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol/issue: 51(1), pp. 107-13, 2008.

[5] Xu Y., *et al.*, "Efficient-means++ Approximation with MapReduce," *Parallel and Distributed Systems, IEEE Transactions on*, vol/issue: 25(12), pp. 3135-44, 2014.

[6] Wu K., *et al.*, "Research and improve on K-means algorithm based on hadoop," in *Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on*, pp. 334-337, 2015.

[7] Chadha A. and Kumar S., "An improved K-Means clustering algorithm: A step forward for removal of dependency on K," in *Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on*, pp. 136-140, 2014.

[8] Anchalia P. P., "Improved MapReduce k-Means Clustering Algorithm with Combiner," in *Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on*, pp. 386-391, 2014.

[9] Mehar A. M., *et al.*, "Determining an optimal value of K in K-means clustering," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pp. 51-55, 2013.

[10] Zhang J., *et al.*, "A 2-tier clustering algorithm with map-reduce," in *ChinaGrid Conference (ChinaGrid), 2010 Fifth Annual*, pp. 160-166, 2010.

[11] Kakooei M. and Shahhoseini H. S., "A parallel k-means clustering initial center selection and dynamic center correction on GPU," in *Electrical Engineering (ICEE), 2014 22nd Iranian Conference on*, pp. 20-25, 2014.

[12] Ghuli P., *et al.*, "A Comprehensive Survey on Centroid Selection Strategies for Distributed K-means Clustering Algorithm," *International Journal of Computer Applications*, vol/issue: 125(5), 2015.

[13] Jain M. and Verma C., "Adapting k-means for Clustering in Big Data," *International Journal of Computer Applications*, vol/issue: 101(1), pp. 19-24, 2014.

[14] Cheung Y. M., "k∗-Means: A new generalized k-means clustering algorithm," *Pattern Recognition Letters*, vol/issue: 24(15), pp. 2883-93, 2003.