International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072

Review of Existing Methods in K-means Clustering Algorithm

Sonu Pandey[1], Lokendra kumar Tiwari[2]

[1]M.Tech Scholar, Department of Computer Science & Engineering, KNIT Sultanpur
[2]Asssitant Professor, Ewing Christian College Allahabad

Abstract - *K-means algorithm is one of the most trendy and important algorithms for data clustering. With this algorithm, data of similar types are tried to be clustered together from a large data set with brute force strategy which is done by repeated calculations. With the advancement in Technology, the data at many domains is generated at higher rates reaching size greater than Petabytes. Significant amount of information is unstructured; semi structured or structured documents spread across the network eg. Images, audio, video, spreadsheets, pdf(s), etc, that contain answer to help us to create new products, refine existing products, improve customer relations. This gave rise to Large set of data and challenges of Big Data which is generally suffered from 3V (Volume, Velocity and Variety) problems. Hadoop is an open source framework designed to overcome 3V challenges. Using Hadoop with K-Means resulted in faster processing of large and complex data set. However, arbitrary preliminary centroids have to be provided in traditional K-Means algorithm. The Convergence to be reach highly depends on the set of preliminary centroids. In this paper we propose a method which takes set of preliminary centroids which has been calculated over Hadoop and afterward run the K-Means algorithm which shows that convergence criteria reach earlier in the most of the cases, hence it will improve efficiency and accuracy of the algorithm.*

Key Words: Data Mining, K-Means clustering, arbitrary preliminary centroids, improved preliminary centroids, Hadoop, MapReduce.

1. INTRODUCTION

With the development and improvement of data mining technology, data clustering algorithm are gradually applied to some fields. The definition of clustering in the academic community can be generalized as follows: first, the similarity of data objects. Data objects within the same cluster have great similarity, but data objects within the different cluster have great non-similarities. Second, the distance of data objects. Take entire data set as a test data object of the gathering, the distance between any pair of data objects within the same cluster size should not be greater than the distance between the different clusters of arbitrary data object. Third, the density of data objects. Take entire data set as a multi-dimensional space aggregation of the data object, a cluster is the spaces which contain the number of data object relatively high dimension cut off by the space which contains the number of data object relatively low dimension. Thus form a relatively separated set of dimensional space.

The k-means algorithms [3, 4 and 11] have been used to produce the clusters with the help of K-Means Algorithm. As we know that traditional K-Means clustering algorithm [4] is mostly dependent of the data set, if the data set is very large it will take more time to go at the convergence stage. Moreover In the most of the cases algorithm results are depends on choice of the arbitrary preliminary centroids. Quite a few attempts have been made by researchers [14] to compute the overall result of the K-Means clustering [11, 13].

In this paper we propose technique to improve Accuracy and Efficiency by producing preliminary Centroids for k-means Clustering over Apache™ Hadoop [6] to harness the power of parallel computing with clustering technique.

1.1 HADOOP-COMPUTATION AND STORAGE SOLUTION

Dealing with "Big Data" requires – an inexpensive, reliable storage and a new tool for analyzing structured, unstructured and semi structured data. Apache Hadoop addresses both of these problems. Because Hadoop works on map reduce concept it share out and parallelize data processing across many nodes in a compute cluster, speeding up large computations and hiding I/O latency through increased concurrency. It is fighting fit for large data processing like searching and indexing in massive data set.

1.2 HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS had been mainly built as transportation for the Apache NUTCH web search engine assignment. HDFS is now an Apache Hadoop subproject. HDFS has master/slave architecture. HDFS is suitable for applications that have large dataset. HDFS maintain the metadata in a dedicated server called NameNode and the application data are kept in separated nodes called DataNode. These server nodes are fully connected and they communicate using TCP based protocol.

2. K-MEANS CLUSTERING OVER HADOOP

The input has been provided to K-Means over Hadoop [10] is given as <key,value> pair, where key is the 'centroid' and 'value' is serialized data nodes(objects) that are need to be clustered. These keys and values are maintained in HDFS in separate files. Centroid file contains preliminary centers either entered by the user or selected arbitrarily from the data nodes(objects) to be clustered. These centers form 'key' for <key, value> pair during Mapper phase.

Operation mechanism of MapReduce is as follows:

i. **Input:** MapReduce framework based on Hadoop requires a pair of Map and Reduce functions implementing the appropriate interface or abstract class, and should also be specified the input and output location and other operating parameters.

ii. **MapReduce:** framework puts the application of the input as a set of key-value pairs <key, value>. In the Map stage, the framework will call the user-defined Map function to process each key-value pairs <key, value>, while generating a new batch of middle key value pairs<key, value>.

iii. **Shuffle :** In order to ensure that the input of Reduce outputted by Map have been sorted, in the Shuffle stage, the framework uses HTTP to get associated key-value pairs <key, value> Map outputs for each Reduce; Map Reduce frame workgroups the input of the Reduce phase according to the key value.

iv. **Reduce :** This phase will traverse the intermediate data for each unique key, and execute user-defined Reduce function. The input parameter is < key, {a list of values} >, the output is the new key-value pairs< key, value >.

v. **Output :** This stage will write the results of the Reduce to the specified output directory location.

3. REVIEW OF DIFFERENT DATA MINING TECHNIQUES

An attempt has been made to study and examine critically all the available findings of previous researches and review the salient features concerned in present work in a well-defined manner and summarized to use it as a background literature in the following paper.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

i. **Pham et al. (2004)** proposed Factors that new measures to assist the selection is proposed and then conclude with an analysis of the results of using the proposed measure to resolve the number of clusters for the k-Means algorithm for dissimilar data sets.

ii. **Fahim et al. (2006)** presented a simple and efficient clustering algorithm based on the k-means algorithm, which they call enhanced k-means algorithm. It is very simple algorithm, which shows the implementation, requiring a simple data structure to keep some information in all iteration to be used in the next iteration. Experimental results demonstrated that scheme can improve the computational speed of the k-Means algorithm by the magnitude in the total number of distance calculations and the overall time of computation.

iii. **Deelers et al. (2007)** gave an algorithm to calculate preliminary cluster centers for k-Means clustering. Data in a cell is partitioned using a cutting plane that divides cell in two smaller cells. The plane is vertical to the data axis with the highest variation and is intended to reduce the sum-squared errors of the two cells as much as possible, while at the same time keep the two cells far apart as possible. Cells have been partitioned one at a time until the number of cells equals to the predefined number of clusters K. The experimental results show that the proposed algorithm is efficient, meet to better clustering results than those of the random initialization. The research also indicated the proposed algorithm would really improve the chances of every cluster containing some data in it.

iv. **Sebastian et al. (2009)** proposed several methods in the literature for improving the performance of the k-Means clustering algorithm. Paper simulates a method for making the algorithm more effective and efficient as to get better clustering with compact complexity.

v. **Chen et al. (2009)** offered a new clustering method based on k-Means that have avoided substitute randomness of initial centre. This work is focused on k-Means algorithm; initial value of the dependence of k selected from the aspects of the algorithm is enhanced. First, the initial cluster number is N. Second, through the application of the sub-merger strategy the category were shared. The algorithm does not require the user to give in advance the number of cluster. Experiments on artificial datasets are presented to have shown considerable improvements in clustering accuracy in association with the random k-Means.

vi. **Pakhira et al. (2009)** presented a modified version of the k-means algorithm that efficiently eliminates the empty cluster difficulty. They described that the updated algorithm is semantically equivalent to the traditional k-Means and there is no performance issue due to integrated modification. Results of simulation experiment using several datasets prove the claim.

vii. **Gupta et al. (2010)** proposed an algorithm to automatically determine the number of clusters in a given input data set, under a combination of Gaussians assumption. The algorithm extends the Anticipation- Maximization clustering approach by preliminary with a single cluster assumption for the data, and recursively split one of the clusters in order to find a tighter fit. An Information standard parameter is used to pick between the present and previous form after each split. The approach is build upon prior work done on both k-Means and Expectation-Maximization algorithms. The algorithm is extended using a cluster splitting approach based on Principal Direction disruptive Partitioning, which improve accuracy and efficiency.

viii. **Yedla et al. (2010)** simulate a new technique for result the improved preliminary centroids and to

provide an efficient way of passing on the data points to appropriate clusters with compact time complexity. According to evaluated results, the modified algorithm has more accurate with less time consuming as compared to original k-means clustering algorithm.

ix. **Ren et al. (2012)** simulate Hadoop workloads from three different clusters on an application-level perspective, with two goals: (a) explore new issues in application patterns and user behavior and (b) understand key performance challenges related to Input/output and load balancing. The carrying out logs from three Hadoop clusters used for research: OPENCLOUD, M45, and WEB MINING. Studied job performance, configurations and user history files from three different Hadoop clusters for academic investigate. These new Hadoop cluster traces contain comfortable information than previous study by recording application pattern and user behaviors, which are critical for understanding the requirements and performance of big-data systems. Easing the use of Hadoop, and improve system designs subject to changing use cases are crucial research information for future.

x. **Dittrich et al. (2012)** states the need of many organizations, companies, and researchers to deal with big data volumes efficiently that includes web analytics applications, scientific applications, and social networks. A trendy data dealing out mechanism for big data is Hadoop MapReduce. Earlier versions of Hadoop MapReduce suffer from various performance problems. There are various strategies that can be used with Hadoop MapReduce jobs to boost up the performance by orders of degree. Jens Dittrich briefly familiarizes the audience with Hadoop MapReduce and motivates its use for big data processing and focuses on different data management techniques, going from job optimization to physical data organization like data layouts and indexes. Through similarities and differences between Hadoop MapReduce and Parallel DBMS are discussed.

xi. **Jain et al. (2012)** proposed a new hybrid algorithm, which is based on k-Means & k-Harmonic Mean approach. Its performance is compare with the customary K-means & K harmonic means algorithm. The outcome which has been obtained from proposed hybrid algorithm is to a great extent better than the traditional K-mean & K harmonic means algorithm.

xii. **Kane et al. (2012)** proposed a new, efficient approach to determine the number of clusters based on the volume of a cluster by comparing it with a fixed threshold.

xiii. **Zhang and Fang (2013)** introduces the idea of the k-means clustering algorithm analysis, the advantages and disadvantages of the traditional k-means clustering algorithm and elaborates the method of improving the k-means clustering algorithm based on improving the initial focal point and thus determine the K value. Experimental results show that the superior clustering algorithm is more stable in clustering process. In the mean time, improved clustering algorithm to reduce or even avoid the impact of the noise data in the dataset object to ensure that the final clustering result is more accurate and effective.

xiv. **Kodinariya and Makwana (2013)** explored six different approaches to determine the right number of clusters in a dataset. There are various methods offered to estimate the number of clusters such as statistical indices, variance based method, Information Theoretic, goodness of fit method etc.

xv. **Anchalia et al. (2013)** discussed the implementation of the K-Means Clustering Algorithm over a distributed environment using Apache™ Hadoop. Here they design the Mapper and Reducer routine for processing of dataset and HDFS has been used for the storage of dataset before and after processing. Mapper takes the input as the <key, value> pair where key work as the centre of the cluster and value is the serializable implementation of the dataset. The initial set of centre has been stored on HDFS prior to the Map function is called and it works as the key for the <key, value> pair. The Mapper is design in such a manner that it computes the distance between the vector value and each of the cluster centers mentioned in the cluster set and simultaneously keeping track of the cluster to which the given vector is closest.

xvi. **Revathi and Nalini (2013)** presented a comparative study of clustering algorithms across two different data items. The result of the variety of clustering algorithms is compared based on the time engaged to form the estimated clusters. Based on experimental results it can be concluded, that the time taken to form the clusters increases as the number of cluster increases. The farthest first clustering algorithm takes very little time to cluster the data items whereas the simple k-Means takes the longest time to perform clustering.

xvii. **Shah Neepa (2014)** discussed the importance of document clustering that emerges from the massive volumes of textual documents created. With more and more development of information technology, data set in many domains is reaching beyond peta-scale; making it difficult to work with the document clustering algorithms in central site and leading to the need of increasing the computational requirements. Parallel computing concepts have been introduced for the elaboration of document clustering which later introduced distributed document clustering. The distributed document clustering using Hadoop and map-reduce has been proposed. First of all k-means has been tested on single node then after modified the mapper and reducer functions to run over cluster of three machines. Dataset consisting of 20,000 documents (20-newsgroups) and 21578 documents were

tested. Results showed that the required time has been reduced that after addition of more nodes.

xviii. **Duggal et al. (2015)** reviews that we live in on-demand, on-command Digital universe with data prolifering by Institutions, Individuals and Machines at a very high rate. This data has been known as "Big Data" due to its sheer Volume, Variety, Velocity and Veracity. The majority of this data are unstructured, structured or semi structured and it is assorted in nature. The degree and the heterogeneity of data are generated with the rapid rate, makes it difficult for the present computing infrastructure to supervise Big Data. Conventional data supervision, warehousing and analysis systems fall short of tools to analyze this data. The authors suggest various methods for catering to the problems in hand through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce is an important method which makes use of file indexing with mapping, sorting, shuffling and finally reducing.

3. CONCLUSION

This paper presents a new and easy technique to generate preliminary set of centroids for Improving the efficiency and accuracy of one dimensional data set. The proposed method fairly reduces the no. of iterations to reach convergence as K-Means is highly sensitive to set of preliminary centroids. The overall execution time for K-Means Clustering job to finish has also reduced. Making the technique handy for large sets of data that would generally require large amount of time to reach convergence as it has been observed in case of arbitrarily selected initial centroids. The result may vary for different data sets.

REFERENCES

[1] Abhijit Kane (2012). Determining the number of Clusters for a K-Means Clustering Algorithm, Indian Journal of Computer Science and Engineering (IJCSE), Vol. 3 No.5 Oct-Nov 2012.

[2] Meenakshi, Poonam Yadav (2016). A survey paper on K-means clustering using hadoop, IJRAET V-4 I-2.

[3] Chunfei Zhang and Zhiyi Fang (2013). An Improved K-means Clustering Algorithm, Journal of Information & Computational Science 10: 1 (2013) 193–199.

[4] J. B. MacQueen (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297.

[5] Jens Dittrich and Jorge Arnulfo Quian´e-Ruiz (2012), Efficient Big Data Processing in Hadoop MapReduce, Very Large Data Bases, Vol. 5, No. 12, 2012.

[6] K. A. Abdul Nazeer and M. P. Sebastian (2009). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009, Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[7] Yashika Verma, Sumit Kumari (2013). Study and analysis on Document Clustering Based on MapReduce in Hadoop using K-mean Algorithm, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[8] Kohei Arai and Ali Ridho Barakbah (2007). Hierarchical K-means: an algorithm for centroids initialization for K-means, Rep. Fac. Sci. Engrg. , Saga Univ. 36-1 (2007), 25-31.

[9] Likas, N. Vlassis and J.J. Verbeek (2003). The Global k-means Clustering algorithm, Pattern Recognition, Volume 36, Issue 2, 2003, pp. 451-461.

[10] P Anchalia, Koudinya, Srinath (2013). MapReduce Design of K-Means Clustering Algorithm.

[11] Fahim A. M., Salem A. M., F.A. Torkey and M.A. Ramadan (2006). An Efficient enhanced k-means clustering algorithm, Journal of Zhejiang University, 10(7): 6261633.

[12] Fang Yuan, Zeng-Hui Meng, Hong-Xia Zhang, Chun-Ru Dong (2004). A New Algorithm To Get The Initial Centroids, Proceedings of the Third International Conference on Machine Laming and Cybernetics, Shanghai, 26-29 August 2004.

[13] S. Deelers, and S. Auwatanamongkol (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance, World Academy of Science, Engineering and Technology Vol:1 2007-11-27.

[14] Revathi and Dr. T. Nalini (2013). Performance Comparison of Various Clustering Algorithm, IJARCSSE, Volume 3, Issue 2, February 2013.