International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume3, Issue 1 (January'2017)

A REVIEW ON BIG DATA ANALYSIS AND K-MEANS CLUSTERING ALGORITHM USING PERFORMANCE OF GPU

K.M.Padmapriya,
Research Scholar,
Bharathiar University,
Coimbatore,Tamilnadu,India.

Dr. B.Anandhi,
Associate Professor and Head,
Department of Computer Science,
Vellalar College for Women,
Erode,Tamilnadu,India.

Abstract: Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. K-means is a simple and widely used algorithm for clustering data. But, the traditional k-means is computationally expensive; sensitive to outlier's i.e. unnecessary data and produces unstable result hence it becomes inefficient when dealing with very large datasets. Solving these Issues is the subject of many recent research works. In this paper, we will do a review on k-means clustering algorithms.

Keywords: Big Data, K-Means Clustering, Map Reduce, Datamining

I. INTRODUCTION

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big data may be important to business – and society – as the Internet has become. Big Data is so large that it's difficult to process using traditional database and software techniques. More data may lead to more accurate analyses. More accurate analyses may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk. Analyzing big data is one of the challenges for researchers system and academicians that needs special analyzing techniques. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data ("big data") to discover patterns and other useful information. Not only will big data analytics help you to understand the information contained within the data, but it will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the *knowledge* that comes from analyzing the data.

Clustering is a common operation that has many applications for data processing [7]. With the rapid growth of digital information collection and analysis, improvements made to this basic workhorse have very significant implications for the operations of large-scale data processing workflows. Internet search engines periodically crawl the web [6], [8], and index the information for fast search and data retrieval. One of the crucial parameters for these services is the liveness of the service, or the time delay before new material is available in the search index. Since the freshness of results is critical to the success of the service, the performance of the underlying data analysis algorithms are essential. Kmeans clustering is one of the fundamental building blocks driving many of the algorithms used by these services. While performance is one factor for the adoption of our approach, efficiency is another leading driver. Data centers consume enormous amounts of energy. Rack space for computers, and inter-node communications present significant spatial, performance and backup electrical generation overhead. If we can offload this processing from the power-hungry CPUs to multiple power efficient GPUs on a single node, significant energy savings can be realized. Likewise, on mobile computers such as notebooks, desktop search, spam filtering [9], antivirus and anomalybased firewall intrusion detection[10], [11], [12] can be greatly accelerated by using the graphics chip to reduce the energy consumed by the main processor. CUDA technology gives computationally intensive applications access to the tremendous processing power of the latest GPUs through a C-like programming interface. The GeForce 8 series GPUs have up to 128 processing elements running at 1.5 GHz and up to 1.5GB of on-board memory. CUDA is designed from the ground-up for efficient general purpose parallel computation on GPUs. It uses a C-like programming language and does not require remapping algorithms to graphics concepts [14].

II.LITERATURE REVIEW

Yugal Kumar, G. Sahoo [1] focused on K-Means initialization problems. The K-Means initialization problem of algorithm is formulated by two ways; first, how many numbers of clusters required for clustering and second, how to initialize initial centers for clusters of K-Means algorithm. This paper covers the solution for of the initialization problem of initial cluster centers. For that, a binary search initialization method is used to initialize the initial cluster points i.e. initial centroid for K-Means algorithm Performance of algorithm evaluated using UCI repository datasets.

Huang Xiuchang, SU Wei [2] focused on problem of user behavior pattern analysis, which has the insensitivity of numerical value, uneven spatial and temporal distribution characteristics strong noise. The traditional clustering algorithm not works properly. This paper analyses the existing clustering methods, trajectory analysis methods, and behavior pattern analysis methods, and combines clustering

algorithm into the trajectory analysis. After modifying the traditional K-MEANS clustering algorithm, the new improved algorithm designed which is suitable to solve the problem of user behavior pattern analysis compared with traditional clustering methods on the basis of the test of the simulation data and actual data, the results shows that the improved algorithm more suitable for solving the trajectory pattern of user behavior problems.

Nidhi Singh, Divakar Singh [3] K-means is widely used for clustering algorithm. This paper proves that the accuracy of k-means for iris dataset is much than the hierarchical clustering and for diabetes dataset accuracy of hierarchal clustering is more than the k-means algorithm. The time taken to cluster the data sets is less in case of k-means. A good clustering method produces high-quality clusters to ensure that objects of a same cluster are more similar than members of different cluster. Kmeans algorithm in this paper works well for large datasets.

Bapusaheb B. Bhusare, S. M. Bansode [4] the K means clustering algorithm which mainly based on initial cluster centers. In this paper K means clustering algorithm by designed in such way that the initial centroids selected using Pillar algorithm. Pillar algorithm effectively chooses the initial centroids and improves accuracy of clusters. However, proposed algorithm has outlier problem leads to reduced performance. So there is need to choose the appropriate parameter in data set for outlier detection mechanism. An improvement in pillar algorithm is done and the number of distance calculation reduced for the previous initial centroids neighbors and used for next step of iterations which causes to increase in the computational time. The experimental results show that the use of pillar algorithm with change improved solution.

Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra [5] found that the K-Means algorithm has two major limitations 1. Several distance calculations of each data point from all the centroids in each iteration. 2. The final clusters depend upon the selection of initial centroids. This work improves k-Means clustering algorithm designed in MATLAB and the datasets from UCI machine learning repository used. The initial centroids initial centroids not selected randomly. By using new approach good clustering results obtained. The new method of selection of initial centroid is better than selecting the initial centroids randomly.

Fahim A. M. et al. [6] proposed an efficient method for assigning data points to clusters. The original k-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original k-means algorithm. Hence there is no guarantee for the accuracy of the final clusters.

Jeyhun Karimov and Murat Ozbayoglu [7] proposed the novel hybrid evolutionary model for the *k*-means clustering algorithm by using metaheuristic methods to identify the good initial centroids for the *k*-means clustering algorithm. The results indicate that the quality of cluster is improved by approximately thirty percent compared to the standard random selection of initial centroids. But still, there is scope for the improvement in the quality. Authors in the paper proposed a method to determine the number of clusters ,k using spectra analysis techniques and then tested this algorithm on various data sets and observed that there is fluctuation in the value of k. Identifying the proper number of clusters from a dataset are two crucial issues in unsupervised learning.

Amira Boukhdhir, Oussama Lachiheb,Mohamed Salah Gouider[8] proposed algorithm an improved KMeans with Map Reduce design for very large dataset. The algorithm takes less execution time as compared to traditional KMeans, PKMeans and Fast KMeans. It removes the outlier from numerical datasets also Map Reduce technique used to select initial centroids and forming the clusters. But it has limitations like the value of numbers of centroids required as input by user. It works on numerical datasets only. Also numbers of clusters are not determined automatically.

In 2009, **Fahim A M et al. [9]** proposed a method to select a good initial solution by partitioning dataset into blocks and applying k-means to each block. But here the time complexity is slightly more. Though the above algorithms can help finding good initial centers for some extent, they are quite complex and some use the k-means algorithm as part of their algorithms, which still need to use the random method for cluster center initialization.

Duong Van Hieu, Phayung Meesad [10] proposed algorithm for reducing executing time of the k-means .They implemented this by cutting off a number of last iterations. In this experiment method 30% of iterations are reduced, so 30% of executing time is reduced, and accuracy is high. However, the choosing randomly the initial centroids produces the instable clusters. Clustering result may be affected by noise points, so it produces inaccurate result.

Li Ma and al [11] developed a solution for improving the quality of traditional k-means clusters. They used the technique of selecting systematically the value of k i.e number of clusters as well as the initial centroids. Also they reduced the number of noise points so the outlier's problem solved. This algorithm produces good quality clusters but it takes more computation time.

Xiaoli Cui and al [12] proposed an algorithm i. e. an improved k-means. This algorithm works on only representative points instead of the whole dataset, using a sampling technique. The result of this the I/O cost and the network cost reduced because of Parallel K-means. Experimental results shows that the algorithm is efficient and it has better performance as compared with k-means but, there is no high accuracy.

Kedar B. Sawan [13] existing K-means clustering algorithm has a number of drawbacks. The selection of initial starting point will have effect on the results of number of clusters formed and their new centroids. Overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional algorithm. The method could be computationally expensive if used with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance. However this drawback could be taken care by using multi-threading technique while implementing it within the program. However further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

Abhijit Kane's [14] paper includes the automatically find the number of clusters in a dataset. Here every step requires re-clustering of the dataset, total O (n) operations computed. This method works well for clusters that are distinctlyseparated. This method is also density-independent, making it useful for clustering algorithms like the Expectation-maximization algorithm.

Omar Kettani, Faical Ramdani, Benaissa Tadili [15] work covers an algorithm designed for automatic clustering. Thismethod computes the correct number of clusters on tested data sets. This method was compared with G-means. The comparison of algorithm shows that the proposed approach much better than G-means in terms of clustering accuracy.

Avni Godara, Varun Sharm [16] covers the prime algorithm. The KMeans clustering is a powerful algorithm used most of the application in daily life dataset, but problem of initial centroid selection. In past years number of papers presented to improve classical k means algorithm. To remove problem of initial centroid selection need to define data points for centroid before next iteration. The use of prim's algorithm gives better results for selection of initial centroid and choose easily data points for future iterations. Experimental result also shows that the prime algorithm gives better and optimal performance for initial centroids, accuracy of result not adjusted.

D. Sharmila Rani, V. T. Shenbagamuthu [17] K-means is a typical clustering algorithm and it is used for clustering largesets of data. This work includes K-means algorithm and analyses the standard K-means clustering algorithm. The standard K-means algorithm is computationally complex and need to reassign the data points, a number of times during every iteration, which makes effect on the efficiency of standard K-means clustering. This paper work covers a simple and efficient way for assigning data points to clusters. This work ensures that the entire process of clustering in O (nk) time without sacrificing the accuracy of clusters.

Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M. [18] Paper build a system to know the accuracy of medication associated with each symptom. To do this K-means Clustering on the clinical note corpus applied. The document clustering results in improving the medication recommendation. An experimental result shows that pre-processing before clustering results in efficient process of clustering. For experimental work different tools used like, section annotator, symptom annotator, negation annotator and medication annotator to get different views of clinical notes which improves the visibility of clinical note. The result of this is increase of the accuracy of medications associated with the symptoms.

Zang B et al., [19] In academia, various parallel data mining algorithms have been proposed on distributed as well as shared memory parallel architecture models in the past decades. However, current available successful commercial mining systems does not applies these parallel data mining techniques, including, IBM Intelligent Miner, SPSS Clementine and SAS Enterprise Miner. Actually neither of these commercial systems achieved real time analysis. One of the reasons that affect the development of parallel data mining algorithms is the high cost of various parallel computing systems. Such as clusters, which are not affordable by small or medium businesses.

T. Zhang, [20] Low power consumption and low cost are the motivating powers of recent development in parallel architectures. One probable recent solution is the multi-core system. Examples are Intel core-duo or core-quad products. Although the multicore systems are consumed low power and low in cost compared with traditional supercomputers. As multiple cores integrated onto a single chip, its scalability is poor. Another popular solution is Graphics Processing Unit (GPU). GPU is originally a highly specialized many-core architecture designed for graphics rendering for the computer gaming industry.

Liheng Jian, [21] Recently high level languages like OpenCL, CUDA (Compute Unified Device Architecture) have developed to support easy programming on GPUs. NVIDIA's GPU with CUDA environment provides standard extensions to C-like languages to manipulate the GPUs. GPUs with CUDA provide tremendous computing power and memory bandwidth for applications. There are computational intensive applications which run on a GPU + CPU heterogeneous system architecture where the GPU acts as the computation accelerator, including scientific, medical, military, business, communication, and other domains. Successful examples are Computational Fluid Dynamics (CFD), Neural Network, Support Vector Machine (SVM), Magnetic Resonance Imaging (MRI), Finite Difference Time Domain (FDTD), intrusion detection, etc.

III.CONCLUSION

Nowadays the processors are mostly multi-core processing. And traditional programming and algorithm are not work efficiency and effective with the hardware. K-means clustering is the most well-known algorithm commonly used for clustering data. In this review work most widely used k-means clustering techniques of data mining is analyzed. This work shows that there are several methods to improve the clustering with different approaches. Various clustering techniques are reviewed which improve the existing algorithm with different perspective.

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 and Technology Vol.62, (2014).

[2]. Huang Xiuchang , SU Wei ,"An Improved K-means Clustering Algorithm" ,JOURNAL OF NETWORKS, VOL. 9, NO. 1, JANUARY 2014

[3]. Nidhi Singh, Divakar Singh," Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012.

[4]. Bapusaheb B. Bhusare, S. M. Bansode, "Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm" , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 4, April 2014 .

[5]. Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra ,"Statistically Refining the Initial Points for K-Means Clustering Algorithm ",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013.

[6]. Fahim A.M,Salem A.M, Torkey A and Ramadan M.A (2006) : An Efficient enchanced k-means clustering algorithm,Journal of Zhejiang University,10(7): 1626-1633,2006.

[7]. Jeyhun Karimov, Murat Ozbayoglu, Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model, *Procedia Computer Science 61 38 – 45* 2015.

[8]. AmiraBoukhdhir Oussama Lachiheb, Mohamed Sala Gouider. "An improved Map Reduce Design of Kmeans for clustering very large datasets", IEEE transaction.

[9]. Fahim A.M,Salem A.M, Torkey F. A., Saake G and Ramadan M.A (2009): An Efficient k-means with good initial starting points, Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19,pp. 47-57.

[10]. V. Duon, M. Phayung. "Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting Method (FMR KMeans)", Springer International Publishing Switzerland, 2015.

[11]. M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, (2015)

[12]. C. Xiaoli and al. "Optimized big data K-means clustering using Map Reduce", Springer Science + Business Media New York (2014).

[13]. Kedar B. Sawant, "Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance "International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015.

[14]. Abhijit Kane," Determining the number of clusters for a Kmeans clustering algorithm", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 3 No.5 Oct-Nov 2012

[15]. Omar Kettani, Faical Ramdani, Benaissa Tadili, "AK-means: An Automatic Clustering Algorithm based on Kmeans ", Journal of Advanced Computer Science & Technology, 4 (2) (2015) .

[16]. Avni Godara, Varun Sharma," Improvement of Initial Centroids in Kmeans clustering Algorithm", Vol-2 Issue-2 2016 IJARIIE

[17]. D. Sharmila Rani, V.T. Shenbagamuthu,"Modified K-Means Algorithm for Initial Centroid Detection", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

[18]. Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M," Enhanced Kmeans clustering approach for healthcare analysis using clinical documents", International Journal of Pharmaceutical and Clinical Research 2016.

[19]. Wu R, Zhang B and Hsu MC,"Cl W.D. Hillis and G.L. Steele Jr., "Data Parallel Algorithms," Comm. ACM, vol. 29, no. 12, pp. 1170-1183, Dec. 1986, doi:10.1145/7902.7903.

[20]. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 103-114, 1996, doi: 10.1145/235968.233324.

[21]. Liheng Jian, ChengWang, Ying Liu, Shenshen Liang,Weidong Yi, Yong Shi, "Parallel data mining techniques on Graphics Processing Unit with Compute Unified Device Architecture(CUDA) ", The Journal of Supercomputing Springer , Volume 64, Issue 3, June 2013.