International Journal of Recent Trends in Engineering & Research
ISSN : 2455-1457 Impact Factor : 3.344

Relevance of K-means Clustering in Big Data Analytics-A Survey

Mugdha Jain[1], Prashant Sharma[2]
[1,2] *Department of Computer Science and Technology,*
Gurukul Institute of Engineering and Technology

Abstract—The growing size of data collected from social networking sites, videos, audios, log files, texts, conversations, documents, medical records, images, tweets, emails, etc have given rise to various issues, the primary being handling such enormous amount of related data while maintaining the time and cost of the operation. This huge amount of data is referred to as Big Data and the task of handling it comes in Big Data Analytics. The various data mining techniques proposed till date serve as an aid to the problem of efficiently analyzing, visualizing and storing Big Data. The K-means clustering algorithm, though proposed more than 50-years ago, serves to be an excellent data mining solution able to cluster this increasing size of data. This paper discusses the various issues encountered in Big Data Analytics over the years and the relevance of the K-means clustering algorithm regarding the same.

Keywords—Big Data; Big Data Analytics; Data mining; Clustering; K-means

I. INTRODUCTION

Big Data refers to the billions and trillions of bytes of information generated from the various activities of humans, whether in the form of digital form or through interactions via the various social networking sites or through textual documents in any business firm or records of the patients all over the world or specifically any area of life which humans are a part of. Around 90% of the total data the world has today is created in no more than the past 2 years. Why Big Data is an issue of concern to almost everybody can be felt from the fact that the size of information humans created till 2003 of around 5exabytes (10^{18} bytes) is reaching equal to the amount generated now by us in just 2 days. 2.72 zettabytes(10^{21} bytes) of digital information calculated in 2012 is expected to double up till now [1]. Handling such large data is now a task for almost all big and small companies and whiles the online and startups are adapting to its size, the technical giants like Google, Facebook, and LinkedIn etc were built expecting Big Data from the beginning. The question concerning each business firm is how to efficiently handle this large and continuously increasing Big data.

Big Data, apart from the notion "BIG" has other various factors that make it a hot research topic for researchers. The data lacks a proper structure and storage, analysis and visualization is still a daunting task for researchers. Apart from this, the data is highly related to each other and consists of diverse sources and new data types which make it all the more complex thereby adding to the high cost and time involved in handling. Handling such data comes in Big Data analytics. Even a small percent of success in reducing the associated complexity is observed to be a huge benefit for any firm when calculated on a large scale [2].

The one feasible solution to analyze Big Data is use clustering, a task of exploratory data mining. Exploring patterns that depict the heterogeneity of data make analysis of data easier. Clustering aims to "cluster" it or group it in a way that the related data fall in separate clusters thereby making searching and sorting easy for such data when need arises. There are many clustering techniques proposed till date, out of which K-means clustering algorithm [3] has proved to be a workhorse in the direction of efficient data mining The reasons attributing to the popularity of K-means [4], even after 50 years of its proposal [5] are the simplicity, scalability and the ease of

implementing the algorithm. Also, it gives a linear asymptotic time complexity with respect to any parameter of the clustering problem. This paper provides a brief survey how K-means can be implemented for Big Data analytics.

II. K-Means Clustering Algorithm

K-means, proposed by E. Forgy [3] in 1965, is a data clustering algorithm aiming to split the given data into k-clusters. It starts with randomly partitioning data points into k-clusters with k-centroids. A new partition is generated through assignment of each point to its closest cluster center or centroid. The third step of the algorithm computes new cluster centers. The assignment step and the third step are repeated until cluster membership is stabilized. The idea behind the proposal was to find a partition with the least Squared Sum Error(SSE) between the empirical mean of cluster and the points that reside in the cluster. Therefore, for K clusters, the SSE can be calculated as

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Where, x_i is the data point in cluster C_k at i[th] dimension and μ_k is the empirical mean of K-clusters and $||x_i - \mu_k||^2$ is the Euclidean distance.
The algorithm has a simple approach to cluster data points, produces scalable results even for large datasets and is easy to implement and use. The algorithm produces spherical clusters because of the use of Euclidean Distance metric. However, the algorithm is encountered with a few problems related to initialization of initial clusters, need of having the prior knowledge of the number of "k" clusters required, generation of only spherically shaped clusters and convergence of the algorithm to local minima. Over the years, many research works have been proposed as improvements to the traditional k-means [3]. The algorithm still produces efficient clustering of large datasets. Applicability of the algorithm for use in Big data is discussed in the next section.

III. APPLICABILITY OF K-MEANS IN BIG DATA

The k-means clustering algorithm as discussed above suffers from a few challenges that fail to make it a fully-efficient clustering algorithm for clustering Big Data, the primary being poor initialization that automatically increases the runtime of the algorithm. Big Data analytics not only require better performance but also a faster convergence. Arthur and Vassilvitski[6] aimed to eliminate the initialization issue of the k-means clustering algorithm and proposed a initialization step different from the conventional k-means algorithm. The first initial center is chosen randomly from the set of data points X and the remaining cluster centers are chosen from the remaining data points of the dataset with a D^2 weighting scheme in an iterative manner where D is the shortest distance of a data point from its nearest cluster center chosen by the algorithm. The rest of the algorithm proceeds according to the conventional k-means. The performance results of the proposed k-means++ algorithm were observed drastically better than the standard k-means. Their proposal was one of the first works in this direction. Some other noteworthy similar research works are done by Bradley and Fayyad [7], Ahmad and Khan [8] and Celebi and Kingravi [9].

Bahmani et al [10] pointed out the sequential nature of the k-means ++ [6] algorithm and the extra number of passes the iterative algorithm uses to converge, which affects the running time of the algorithm. A more scalable approach, as proposed by the authors, by the name Scalable k-means ++, aims to drop down the number of k passes by avoiding the previous criterion of considering every single data point and rather consider a sample number of data points thereby reducing the passes of the algorithm. Scalable k-means is therefore, a parallel version of the k-means++ algorithm [6].

Shindler et al [11] in 2011 used k-means algorithm for streaming data model taking in as input few portion of the data at a time and continuing sequentially in "phases", as in a hard disk file with no random access to data in between. The authors further improve the work by [12] who proposed running the online facility location algorithm of [13] through "guessing" the optimum cost of the algorithm and ending a "phase" when the cost exceeds the pre-determined "guess" or the facilities exceeds some predefined constant "K". The algorithm [12] proceeds as the k-means methodology resulting in a large approximation factor and running time. The improvements by Shindler et al [11] include the removal of unnecessary checks in the previous algorithm, changes in the end of phase criterion and resultant faster worst-case runtime approximation factor and runtime.

To compensate the heterogeneity of data, Cai et al [14] proposed Multi-view K-means clustering on Big Data. Multi-view clustering aims to integrate different views of data that depict different perspectives of the data or the heterogeneous sources the data has arrived from. All the prior similar works are limited to clustering through the graph based approaches found unable to handle large scale data. Also, they incur an additional overhead cause by construction of graphs and eigen decomposition. The authors applied the same to K-means clustering using parallel processing of data and performing clustering on multi-core processors. The proposed algorithm is robust in terms of handling outliers and in anyway, does not hamper the performance of the standard k-means algorithm.

Li et al [15] proposed eliminating the issues of the traditional k-means clustering algorithm for clustering Big Data. The k-means clustering algorithm, being efficiently able to handle the increasing size of data brings with it an increased time complexity. Authors proposed optimizing k-means according to the Hadoop cloud computing platform and MapReduce Framework that allows distributed and parallel processing of data. The algorithm starts by initialization of the cluster centers followed by partitioning the dataset into equally sized small data blocks for parallel processing. The blocks are then exposed to the Map and Reduce tasks that run till the desired clustering results are achieved. Optimization of the k-means algorithm is also done in terms of initialization of cluster centers that otherwise cause instability in clustering results.

Feldman et al [16] proposed using coresets of large data instead of using large dataset for clustering purposes. A coreset can be defined as a subset of the dataset with same properties of the dataset. Running the clustering algorithm on a coreset helps in cutting down the query processing time though satisfying the exact constraints and optimality definitions as used by the dataset. The proposal relaxes the knowledge boundations of the previous algorithms of knowing the number of data points and the dimensions in advance and is limited to taking them in increasing order for each new inserted value. Using the merge and reduce paradigms, the k-means, PCA and projected clustering are made into parallel streaming clustering algorithms.

Eren et al [17] in 2015 proposed using K-means for extracting information from a "Reality Commons" dataset of the MIT Human Dynamics Laboratory consisting of information regarding the various communities with around 100 people in each community. The dataset contains data about the various activities a student's life revolves around in a dormitory. The purpose of the proposal is to deduce how students' eating habits relate to them getting cold. For the vast dataset considered, a parallel partitioning technique, Map Reduce is used on the popular Apache Hadoop framework.

IV. CONCLUSION

Big Data is a topic of great relevance in almost every field. Apart from the three V's of Volume, Velocity and Variety, the term lacks a proper structure and an increased complexity in handling such enormous data size. Data mining allows studying patterns of data to further allow processing on it. Efficiently clustering the increasing size of data to analyze, store, query, transfer

and search for it at any time instance can be done through the various clustering techniques proposed till date. Among all, the k-means clustering algorithm because of simplicity, ease of use, coping ability to cluster increasing size of data through both sequential and parallel approach make it a perfect solution for the Big Data clustering problem. This paper provides a brief overview of some of the noteworthy research works in this direction.

REFERENCES

i. Intel IT Center, "Peer Research: Big Data Analytics", Intel's IT Manager Survey on How Organizations Are Using Big Data, August 2012.
ii. Big Data in Big Companies, Research Report by Tom Davenport and Jill Dyché
iii. E.W.Forgy, "Cluster analysis of multivariate data: efficiency v/s interpretability of classifications", Biometrics, 21, pp. 768–769, 1965
iv. X. Wu et al. "Top 10 algorithms in data mining", Knowledge and Information Systems, 14(1):1{37, 2008}
v. Anil K. Jain, "Data Clustering: 50 Years Beyond K-Means", Department of Computer Science & Engineering Michigan State University East Lansing, Michigan 48824 USA
vi. D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", Society for Industrial and Applied Mathematics , Philadelphia, 2007.
vii. P. S. Bradley and U. Fayyad, "Refining initial points for k-means clustering", Proceedings of the 15th International Conference on Machine Learning, pp. 91–99, 1998.
viii. S.S. Khan and A. Ahmad "Cluster center initialization algorithm for kmeans clustering" Pattern Recognition Letters, Vol. 25, Issue 11, pp. 1293–1302, 2004.
ix. M.E. Celebi and H.A. Kingravi, "Deterministic Initialization of the K-Means Algorithm using Hierarchical Clustering" International Journal of Pattern Recognition and Artificial Intelligence, Vol. 26, Issue 7, pp. 1250018-1250034, November 2012.
x. Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar and Sergei Vassilvitskii, "Scalable k-means", Proceedings of the VLDB Endowment, Vol. 5, No. 7, pp. 622-633, 2012.
xi. Michael Shindler, Alex Wong and Adam W Meyerson, "Fast and accurate k-means for large datasets", Advances in neural information processing systems, pp. 2375-2383, 2011.
xii. Vladimir Braverman, Adam Meyerson, Rafail Ostrovsky, Alan Roytman, Michael Shindler,and Brian Tagiku, " Streaming k-means on Well-Clusterable Data", In SODA, 2011.
xiii. Adam Meyerson, "Online facility location", In FOCS, 200l.
xiv. Xiao Cai, Feiping Nie and Heng Huang, "Multi-View K -Means Clustering on Big Data", Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2598-2604, 2013.
xv. Zhihua Li, Xudong Song, Wenhui Zhu and Yanxia Chen, "K-means Clustering Optimization Algorithm Based on MapReduce ", Proceedings of the International Symposium on Computers & Informatics (ISCI 2015), pp. 198-203, 2015.
xvi. Dan Feldman, Melanie Schmidt, Christian Sohler, "Turning Big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering", Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1434-1453, 2013.
xvii. Beste Eren, Ezgi Çılga Karabulut, S. Emre Alptekin and Gülfem Işıklar Alptekin, "A K-Means Algorithm Application on Big Data", Proceedings of the World Congress on Engineering and Computer Science (WCECS 2015),Vol II, 2015. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

